A RANDOM FOREST APPROACH FOR SOIL MOISTURE ESTIMATION AT 60 METERS SPATIAL RESOLUTION

Gerard Portal[1], Mercè Vall-llossera[1,2], Carlos López-Martínez[1,2], Adriano Camps[1,2,3], Miriam Pablos[2], David Chaparro[4], Amir Mustofa Irawan[2], Alberto Alonso-González[2], Thomas Jagdhuber[4,5]

[1]Institut d'Estudis Espacials de Catalunya – IEEC, 08034 Barcelona, Spain
[2]CommSensLab – UPC, Department of Signal Theory and Communications, Universitat Politècnica de Catalunya (UPC) and IEEC-UPC, 08034 Barcelona
[3]ASPIRE Visiting International Professor, UAE University CoE, 15551 Al Ain, UAE
[4]German Aerospace Center, Microwaves and Radar Institute, 82234 Wessling, Germany
[5]University of Augsburg, Institute of Geography, 86159 Augsburg, Germany

ABSTRACT

A Random Forest (RF) regression-tree method to derive high-resolution (60 m) surface soil moisture maps is proposed in this study. The developed methodology integrates multi-source synergies by incorporating information from the visible, near-infrared until short-wave infrared spectrum (Sentinel-2), reanalysis data (ERA5-Land) and terrain information (SRTM), using exclusively open access data. The analysis focuses on the central part of the Iberian Peninsula and covers a four-year period (2018-2021). The resulting high-resolution soil moisture maps exhibit greater spatial heterogeneity compared to the ESA Climate Change Initiative (CCI) soil moisture, which was used as a reference in the training of the RF model. These maps have been evaluated using *in situ* soil moisture measurements from the REMEDHUS network, and show good agreement in terms of Pearson's correlation (0.83), and uRMSE (0.028 $m^3 \cdot m^{-3}$), demonstrating the method's significant potential for deriving high-resolution soil moisture information.

Index Terms— Soil moisture, random forest, multispectral instrument, Sentinel-2, ESA CCI.

1. INTRODUCTION

In the current context of climate change, extreme weather events are becoming more frequent and understanding Soil Moisture (SM) dynamics is of paramount importance.

Microwave remote sensing, including passive and active acquisition techniques, has enabled the estimation of SM from a regional to a global scale [1–3]. Microwave radiometers exhibit a high radiometric sensitivity (SM accuracy ~0.04 m^3/m^3) and frequent revisit rates (1-3 days). However, their spatial resolutions are limited to tens of kilometers due to physical and technical constraints, such as antenna size. On the other hand, microwave radars offer a spatial resolution in the range of a few meters. Nevertheless, their backscatter measurements may be more easily influenced by vegetation canopy and soil roughness, and are limited to a temporal resolution of about a week or longer.

An increasing number of applications requiring high spatial accuracy (< 1 km) has spurred the development of pixel disaggregation techniques aimed at enhancing the spatial resolution of the traditional radar/radiometer-based SM maps [4–6]. Currently, the explosive growth of multi-sensor and multi-resolution Earth Observation (EO) data, coupled with the significant advancements in statistical learning, has led to the development of a plethora of Machine Learning (ML) downscaling approaches [7–9].

ML algorithms have the capability to leverage heterogeneous information and identify complex nonlinear relationships directly and only from data. Nevertheless, their implementation for high-resolution SM estimation requires careful consideration of several key factors. First, a large number of samples to draw accurate predictions and generalize effectively is needed. Second, it is essential to identify the most relevant variables that influence SM and discard those with little impact, reducing noise in the data. Feature selection techniques can be employed. Third, ML algorithms can be categorized in supervised and unsupervised, depending on the presence of labeled training data. Among the supervised algorithms, regression-trees methods have demonstrated their robustness in previous SM downscaling studies [10], [11], along with the utilization of cross-validation techniques to prevent overfitting. Due to all aforementioned reasons, this study proposes utilizing the Random Forest (RF) regression-tree method [12] for estimating SM at a spatial resolution of 60 m.

2. STUDY AREA & DATA DESCRIPTION

The selected study region corresponds to the central part of the Iberian Peninsula, extending longitudinally from 7.2°W to 3.5°W, and latitudinally from 38.9°N to 42.5°N (Fig. 1a). It is characterized by a Mediterranean-continental and

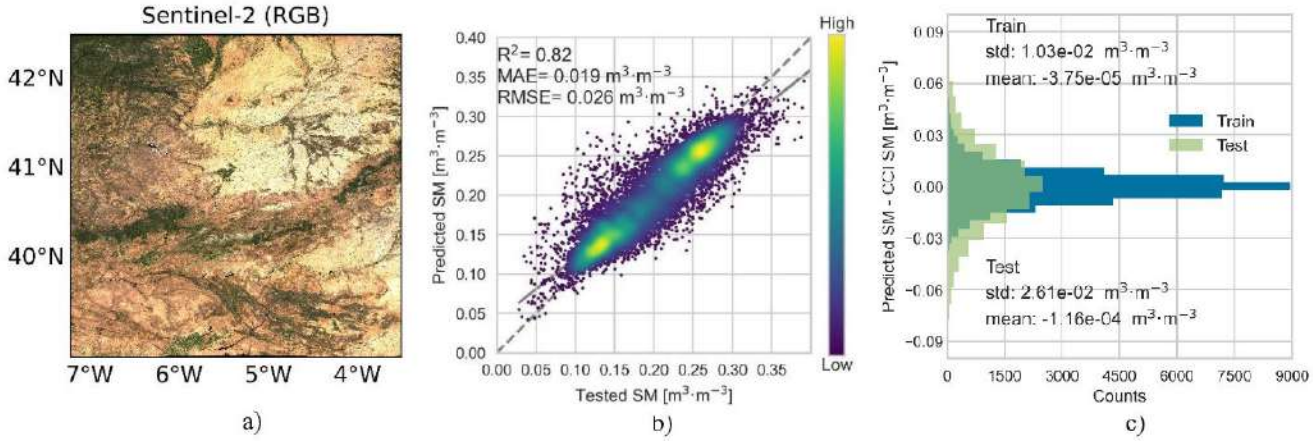

Figure 1. a) Sentinel-2 true color image (with gain and gamma corrections) of the study region, utilizing all available samples from August 2019. b) Comparison between the CCI and the predicted SM (both at 0.25°), using exclusively testing data from 2018 to 2021.The statistics R^2, MAE and RMSE are included. c) Error obtained between the CCI and the predicted SM, using data from 2018 to 2021. The training (blue) and testing (green) phases are depicted. The mean and the std are included for both.

mountainous climate. The study period spans from January 2018 to December 2021. The proposed approach utilizes the following open access datasets (Table 1):

1) Daily ERA5-Land [13] Skin Temperature (SkT) at 12:00 UTC, at a 9 km grid.
2) Sentinel-2 [14] images captured by its MultiSpectral Instrument (MSI), including 11 bands covering the visible and near-infrared to the short-wave infrared spectrum (excluding bands 1 and 10) and 10 vegetation and soil indices, at 10/20/60 m resolutions. The indices used are: the Normalized Difference Vegetation Index (NDVI), the Normalized Difference Red Edge Index (NDRE), the Enhanced Vegetation Index (EVI), the Green Normalized Difference Vegetation Index (GNDVI), the Soil Adjusted Vegetation Index (SAVI), the Normalized Difference Moisture Index (NDMI), the Moisture Stress Index (MSI), the Normalized Burned Ratio Index (NBRI), the Bare Soil Index (BSI), and the Normalized Difference Water Index (NDWI). Sentinel-2 has a revisit frequency of 5 days considering the combined constellation. Due to its revisit period and the potential masking of information by atmospheric effects, particularly in the visible range, datasets derived from Sentinel-2 are the most limiting variables, in terms of spatio-temporal coverage, among all the predictors used in the ML algorithm.
3) Datasets that describe the terrain features, i.e., the Digital Elevation Model (DEM) from the Shuttle Radar Topography Mission (SRTM) [15], and the slope calculated from the DEM.
4) Daily Climate Change Initiative (CCI) SM [13] data at a spatial resolution of 0.25°, derived from the combination of active and passive microwave observations.
5) Daily average of the SM data at the topsoil 5 cm from the Soil Moisture Measurements Stations Network of the University of Salamanca (REMEDHUS) [16], located in the central part of the Duero Basin, in Spain, and consisting of 19 stations (available within the study period) equipped with Hydra Probes, and 4 meteorological stations.

Table 1. Summary of the variables used in this study.

Source	Variable	Resolution	Frequency
Sentinel-2	11 reflectances	10/20/60 m	5 days
	10 indices	10/20 m	5 days
ERA5-Land	Skin temperature	9 km	daily
SRTM	DEM	30 m	Static
	Slope	30 m	Static
CCI	Soil moisture	~25 km	daily
REMEDHUS	Soil moisture	*in situ*	hourly

3. METHODOLOGY

3.1. Model implementation

The procedure conducted to estimate SM at a 60 m spatial resolution is summarized as follows:

1) Preprocess the Sentinel-2 images by filtering out defective pixels, water-covered areas and regions affected by clouds. This information is then aggregated to a regular 60 m grid.
2) Aggregate all the predictors (Sentinel-2 data, ERA5-Land SkT, and terrain data) to the resolution of the target variable (0.25° in our case) by employing the median.
3) Apply the RF algorithm at low resolution (0.25°) to establish the relationship between the predictors and the target variable (CCI SM). The data are randomly divided into training (75%) and testing (25%) subsets. The 'RandomForestRegressor' class, available within the 'scikit-learn' open-source ML library for Python [17], is

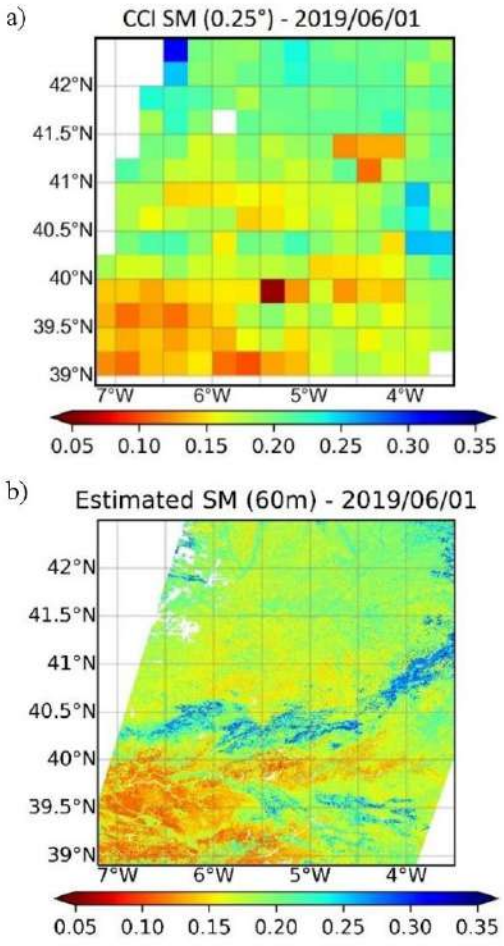

Figure 2. a) CCI SM map on a 0.25° grid and b) estimated SM at 60 m, both for June 1st, 2019.

used to build a model for predicting SM. Hyperparameter tuning of the RF class is performed using a k-fold (k=5) cross-validation technique.

4) Generate 60 m SM maps using all the predictors at high resolution – Sentinel-2-related data (reflectances and indices) are already at 60 m, the ERA5-Land SkT is linearly interpolated to the 60 m, grid and the terrain data is aggregated from 30 m to 60 m – using the regression model obtained in step 3.

3.2. Temporal validation of the 60 m SM

The original CCI SM at 0.25° and the resulting 60 m SM were compared with the *in situ* SM measurements provided by the stations of REMEDHUS. First, the daily average of the most representative stations (which are F11, H13, J12, J14, K10, M9 and O7 according to [11]) was obtained. The same methodology was applied to calculate the daily average of the CCI SM and the 60 m pixels containing these seven stations. A statistical analysis was then conducted using an equal number of samples.

4. RESULTS

The analysis reveals that the seven most significant predictors (and their contribution scores) during the ML training are: the skin temperature (47.73%); the month of the year (12.24%); the Short-Wave InfraRed (SWIR) band 11 (6.6%); the DEM (6.2%); the slope (3.23%); the SWIR band 12 (2.1%); and the BSI (1.8%). Fig. 1b illustrates the relationship between the CCI SM (0.25°) and the predicted SM (0.25°), yielding an explained variance score (R^2) of 0.82, a Mean Absolute Error (MAE) of 0.019 $m^3 \cdot m^{-3}$, and a Root Mean Square Error (RMSE) of 0.026 $m^3 \cdot m^{-3}$. Fig. 1c displays the histogram of the errors (Predicted SM minus CCI SM), demonstrating a similar Gaussian pattern centered around zero $m^3 \cdot m^{-3}$ for both the training and the testing data, with a slightly higher standard deviation (std) for the testing data errors. Fig. 2 presents an example (June 1st, 2019) showing the original CCI SM (Fig. 2a) and the estimated SM at high resolution (Fig. 2b), where an increase in spatial heterogeneity can be observed. Temporally, these two products were compared with the *in situ* SM measurements provided by the REMEDHUS network stations (Fig. 3). They exhibit good agreement with the *in situ* measurements, with slightly higher correlation observed for the CCI SM product, but also showing a greater bias. Both products have an unbiased Root Mean Square Error (uRMSE) of 0.028 $m^3 \cdot m^{-3}$, lower than the nominal SM accuracy usually required in satellite missions (0.04 $m^3 \cdot m^{-3}$).

5. CONCLUSIONS

This study introduces a RF technique to estimate high-resolution SM maps, accounting for multi-source synergies. In the specific implementation presented here, the skin temperature (which has proved to be highly relevant in other disaggregation techniques [4]), the month of the year, and the Sentinel 2 SWIR band emerge as the most significant predictors, surpassing other bands or indices (e.g., NDVI or NDRE). The initial results are promising, with an R^2 value of 0.82 between the CCI SM and the estimated SM (0.25°, Fig. 1b). The model successfully captures the complex spatial heterogeneity of the terrain at 60 m resolution for the Iberian Peninsula.

The algorithm encounters some difficulties in predicting rare extreme events (Fig. 2a; dark blue pixel, 42.3°N 6.3°W; dark red pixel, 39.8°N 5.3°W) where SM is above ~0.3 $m^3 \cdot m^{-3}$ or below ~0.1 $m^3 \cdot m^{-3}$. This limitation is likely due to the scarcity of samples within these ranges during the training phase. One potential solution could be to expand the study area to include wetter and drier regions, such as the northern and northwestern parts, as well as the southern region of Spain, respectively.

The time series presented in Fig. 3 demonstrates that the high-resolution estimated SM effectively captures dry-down and rewetting events. However, there is a noticeable bias (0.074 $m^3 \cdot m^{-3}$) compared to the *in situ* measurements, which

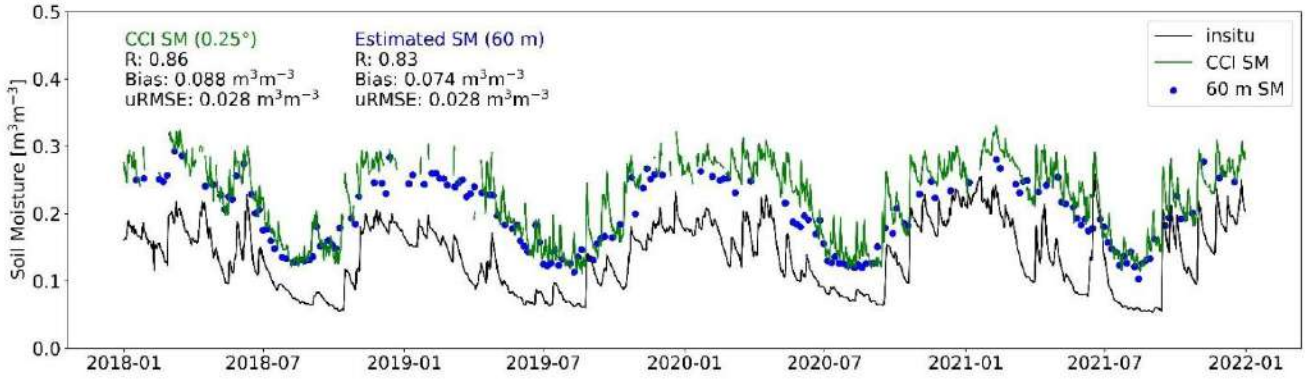

Figure 3. Comparison between the original CCI SM (green line), the estimated SM at 60 m (blue dots), and the *in situ* measurements (black line).

can be attributed to inherent biases present in the reference data used for model training or differences in spatial resolutions (*in situ* vs. 60 m resolution). The time series also reveals a limited number of samples for the 60 m estimated SM, mainly due to the temporal synchronism required between the Sentinel-2 data and the CCI SM target variable, as well as clouds masking the optical information.

8. ACKNOWLEDGMENTS

This research was supported by the Spanish Ministry of Science and Innovation (MCIN/AEI/10.13039/501100011033), through the project INTERACT "PID2020-114623RB-C32. This work was also supported by the Institute d'Estudis Espacials de Catalunya - IEEC. David Chaparro received funding from the Fundación Ramón Areces. In addition, Amir M. Irawan received the support of a fellowship from "La Caixa" Foundation (ID 100010434), with the fellowship code LCF/BQ/DI21/11860028. The authors would like to thank the Water Resources Research group of the University of Salamanca for their support.

9. REFERENCES

[1] T. Schmugge, «Applications of passive microwave observations of surface soil moisture», *Journal of Hydrology*, vol. 212-213, pp. 188-197, dic. 1998.

[2] Y. H. Kerr *et al.*, «The SMOS Mission: New Tool for Monitoring Key Elements ofthe Global Water Cycle», *Proceedings of the IEEE*, vol. 98, n.° 5, pp. 666-687, may 2010.

[3] D. Entekhabi *et al.*, «The Soil Moisture Active Passive (SMAP) Mission», *Proceedings of the IEEE*, vol. 98, n.° 5, pp. 704-716, may 2010.

[4] G. Portal *et al.*, «A Spatially Consistent Downscaling Approach for SMOS Using an Adaptive Moving Window», *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, n.° 6, pp. 1883-1894, jun. 2018.

[5] N. N. Das *et al.*, «The SMAP and Copernicus Sentinel 1A/B microwave active-passive high resolution surface soil moisture product», *Remote Sensing of Environment*, vol. 233, p. 111380, nov. 2019.

[6] G. Portal *et al.*, «Impact of Incidence Angle Diversity on SMOS and Sentinel-1 Soil Moisture Retrievals at Coarse and Fine Scales», *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-18, 2022.

[7] S. O. y R. Orth, «Global soil moisture data derived through machine learning trained with in-situ measurements», *Sci Data*, vol. 8, n.° 1, jul. 2021.

[8] J. Du, J. S. Kimball, R. Bindlish, J. P. Walker, y J. D. Watts, «Local Scale (3-m) Soil Moisture Mapping Using SMAP and Planet SuperDove», *Remote Sensing*, vol. 14, n.° 15, Art. n.° 15, ene. 2022.

[9] N. Vergopolan *et al.*, «SMAP-HydroBlocks, a 30-m satellite-based soil moisture dataset for the conterminous US», *Sci Data*, vol. 8, n.° 1, Art. n.° 1, oct. 2021.

[10] Z. Wei, Y. Meng, W. Zhang, J. Peng, y L. Meng, «Downscaling SMAP soil moisture estimation with gradient boosting decision tree regression over the Tibetan Plateau», *Remote Sensing of Environment*, vol. 225, pp. 30-44, may 2019.

[11] G. Portal *et al.*, «Assessment of Multi-Scale SMOS and SMAP Soil Moisture Products across the Iberian Peninsula», *Remote Sensing*, vol. 12, n.° 3, ene. 2020.

[12] L. Breiman, «Random Forests», *Machine Learning*, vol. 45, n.° 1, pp. 5-32, oct. 2001, doi: 10.1023/A:1010933404324.

[13] J. Muñoz-Sabater *et al.*, «ERA5-Land: a state-of-the-art global reanalysis dataset for land applications», *Earth System Science Data*, vol. 13, n.° 9, pp. 4349-4383, sep. 2021.

[14] Sentinel-2 PDGS Project Team, «Sentinel-2 Calibration and Validation Plan for the Operational Phase». 22 de diciembre de 2014.

[15] T. G. Farr *et al.*, «The Shuttle Radar Topography Mission», *Reviews of Geophysics*, vol. 45, n.° 2, 2007.

[16] N. Sanchez, J. Martinez-Fernandez, A. Scaini, y C. Perez-Gutierrez, «Validation of the SMOS L2 Soil Moisture Data in the REMEDHUS Network (Spain)», *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, n.° 5, pp. 1602-1611, may 2012.

[17] F. Pedregosa *et al.*, «Scikit-learn: Machine Learning in Python», *MACHINE LEARNING IN PYTHON*.