Supplementary material

Live fuel moisture content mapping in the Mediterranean basin using Random Forests and combining MODIS spectral and thermal data

Àngel Cunill Camprubí, Pablo González-Moreno, Víctor Resco de Dios

S1. Land surface temperature

We performed preliminary tests to specify the land surface temperature (LST) product that would be used in the development of the model. The comparison was made using the forward feature selection (FFS) process described in section 2.2.3. Specifically, we tested the ability of the MODIS LST daily (MOD11A1) and the 8-day average composite (MOD11A2) products to predict LFMC along with the other variables in a variable selection process, where its usefulness is also evaluated. MOD11A1 data were obtained for each sample site by their sampling date. MOD11A2 were extracted from the composite layer which includes, on their averaged days, the corresponding sampling date. The number of missing data was much greater in MOD11A1 than in MOD11A2 (34.7% and 7.7% after removing missing values from reflectance data, respectively). In order to be a fair comparison, we formed a single dataset of equal size by eliminating missing values in both variables. The two variables were selected in the FFS. The final models reached an RMSE of 20.38% and 20.27% for the daily LST and 8-day composite product, respectively. Given that RMSEs were very similar between both models, and that the 8-days LST composite showed smaller data gaps, we used the latter.

S2. Data extraction method

We previously tested the method for remote-sensing data extraction at the sampling sites. In particular, we compared the performance of models with all predictor variables obtained from a simple pixel extraction or the average value from the 3×3 pixels window closest to each field sampling location. To do so, we used a nested 5-fold leave-location-out cross-validation and 30 iterations, as we describe in section 2.2.4. Average mean bias error (MBE), root mean square error (RMSE) and variance explained by cross-validation (VE_{CV}) are shown in Table S1. Additionally, we applied a Chi-square test using the distributions of RMSE and VE_{CV} values from the 30 repetitions to statistically check differences. We concluded that there were no significant differences between both

methods (RMSE p-value = 0.9; VE_{CV} p-value = 0.88). Averaging of adjacent pixels to avoid positioning error may be considered advantageous at higher spatial resolutions (e.g., Sentinel-2) than those treated here (Congalton & Green, 2019) and we used simple pixel extraction for simplicity and shortest computation time.

Table S1. Performance metrics from the focal mean and simple pixel extraction comparison.

Method	**MBE [%]**	**RMSE [%]**	**VE_{CV}**
Focal mean	1.33	20.52	0.33
Simple pixel	1.05	20.53	0.33

S3. Spectral vegetation indices

Table S2. Spectral vegetation indices used to estimate LFMC based on the MCD43A4 Collection 6 reflectance bands: B1, Red; B2, NIR1; B3, Blue; B4, Green; B5, NIR2; B6, SWIR1; B7, SWIR2. MODIS formulations extracted from the literature cited in the main text.

Index	**Formulation**	**Reference**
Normalized Difference Vegetation Index	$NDVI = \frac{B2 - B1}{B2 + B1}$	Rouse et al. (1974)
Normalized Difference Water Index	$NDWI = \frac{B2 - B5}{B2 + B5}$	Gao (1996)
Normalized Difference Infrared Index	$NDII6 = \frac{B2 - B6}{B2 + B6}$	Hardisky et al. (1983)
Normalized Difference Infrared Index (with band 7)	$NDII7 = \frac{B2 - B7}{B2 + B7}$	Hardisky et al. (1983)
Global Vegetation Moisture Index	$GVMI = \frac{(B2 + 0.1) - (B6 + 0.02)}{(B2 + 0.1) + (B6 + 0.02)}$	Ceccato et al. (2002)
Enhanced Vegetation Index	$EVI = \frac{2.5 \times (B2 - B1)}{B2 + 6 \times B1 - 7.5 \times B3 + 1}$	Huete et al. (2002)
Soil Adjusted Vegetation Index	$SAVI = \frac{(1 + 0.5)(B2 - B1)}{B2 + B1 + 0.5}$	Huete (1988)
Visible Atmospherically Resistant Index	$VARI = \frac{B4 - B1}{B4 + B1 - B3}$	Gitelson et al. (2002)
Vegetation Index — Green	$VI_{green} = \frac{B4 - B1}{B4 + B1}$	Tucker (1979)
Normalized Difference Tillage Index	$NDTI = \frac{B6 - B7}{B6 + B7}$	van Deventer et al. (1997)
Simple Tillage Index	$STI = B6/B7$	van Deventer et al. (1997)
Moisture Stress Index	$MSI = B6/B2$	Rock et al. (1986)
Greenness index	$G_{ratio} = B4/B1$	Zarco-Tejeda et al. (2005)

S4. Land cover definitions

Table S3. Land cover classes from samples used in the study. International Geosphere-Biosphere Programme (IGBP) definitions and corresponding grouped classes for the analyses.

IGBP class	Definition	Grouped class
Evergreen needleleaf forests	Dominated by evergreen conifer trees (canopy >2 m). Tree cover >60%.	Forests
Evergreen broadleaf forests	Dominated by evergreen broadleaf and palmate trees (canopy >2 m). Tree cover >60%.	Forests
Mixed forests	Dominated by neither deciduous nor evergreen (40-60% of each) tree type (canopy >2 m). Tree cover >60%.	Forests
Woody savannas	Tree cover 30-60% (canopy >2 m).	Savannas
Savannas	Tree cover 10-30% (canopy >2 m).	Savannas
Open shrublands	Dominated by woody perennials (1-2 m height) 10-60% cover.	Shrublands
Closed shrublands	Dominated by woody perennials (1-2 m height) >60% cover.	Shrublands
Grasslands	Dominated by herbaceous annuals (<2 m).	Grasslands

S5. Model parametrization

The Random Forest algorithm requires specification of some hyperparameters for model calibration. The parameters considered here (Table S4) were the number of variables randomly selected at each split (*mtry*), the total number of trees in the forest (*ntree*), the minimal terminal node size (*min. node size*), and the ratio of observations sampled for each decision tree (*sample fraction*).

Table S4. Boundaries of the RF hyperparameters grid-search space, adjusted parameters for the Forward Feature Selection (FFS) process and optimized hyperparameters for the final model.

Step	Type	$NDVI_{CV}$	ntree	mtry	min. node size	sample fraction
Grid-Search	Start	0.20	250	2	5	0.2
	End	0.60	1000	$p \times 0.40$ p min = 4	30	0.95
	Step	0.05	250	1	1	0.05
FFS	Selected	-	250	2	5	0.632*
CAL	Optimal	-	500	2	5	0.3

*Sample with replacement.; *p*, predictor variables.

The samples selection was made without replacement except for the Forward Feature Selection process. In this case, *mtry* was set to 2, as suggested by Meyer et al. (2019). The number of trees was fixed to 250 to reduce the computational time, as we showed no increase of performance by using more trees. The rest of the parameters were left as configured by default in the RF algorithm.

S6. Data description and features correlation

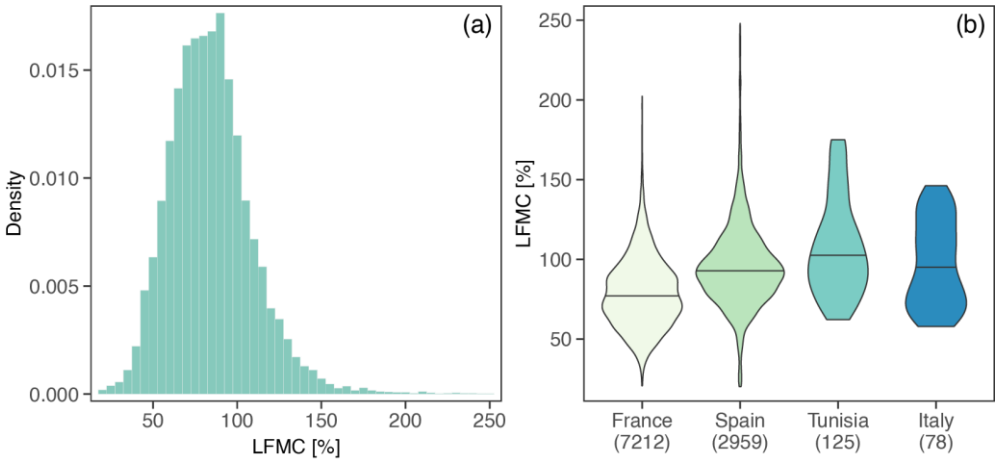

Fig. S1. LFMC ground samples overall (a) and by country (b) distributions. Numbers between parenthesis under country names are the number of samples in that country.

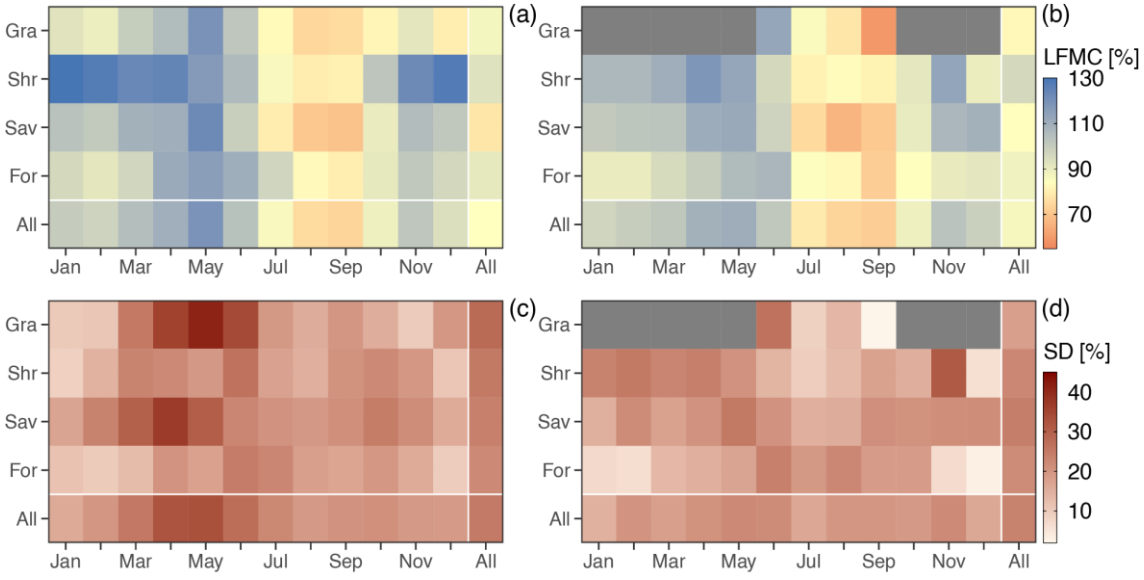

Fig. S2. Mean and standard deviation (SD) matrices from CAL (a, c) and EXT (b, d) of the LFMC field measurements shown by fuel type and month of the year, and the overall of each one. Gray cells indicate no data availability.

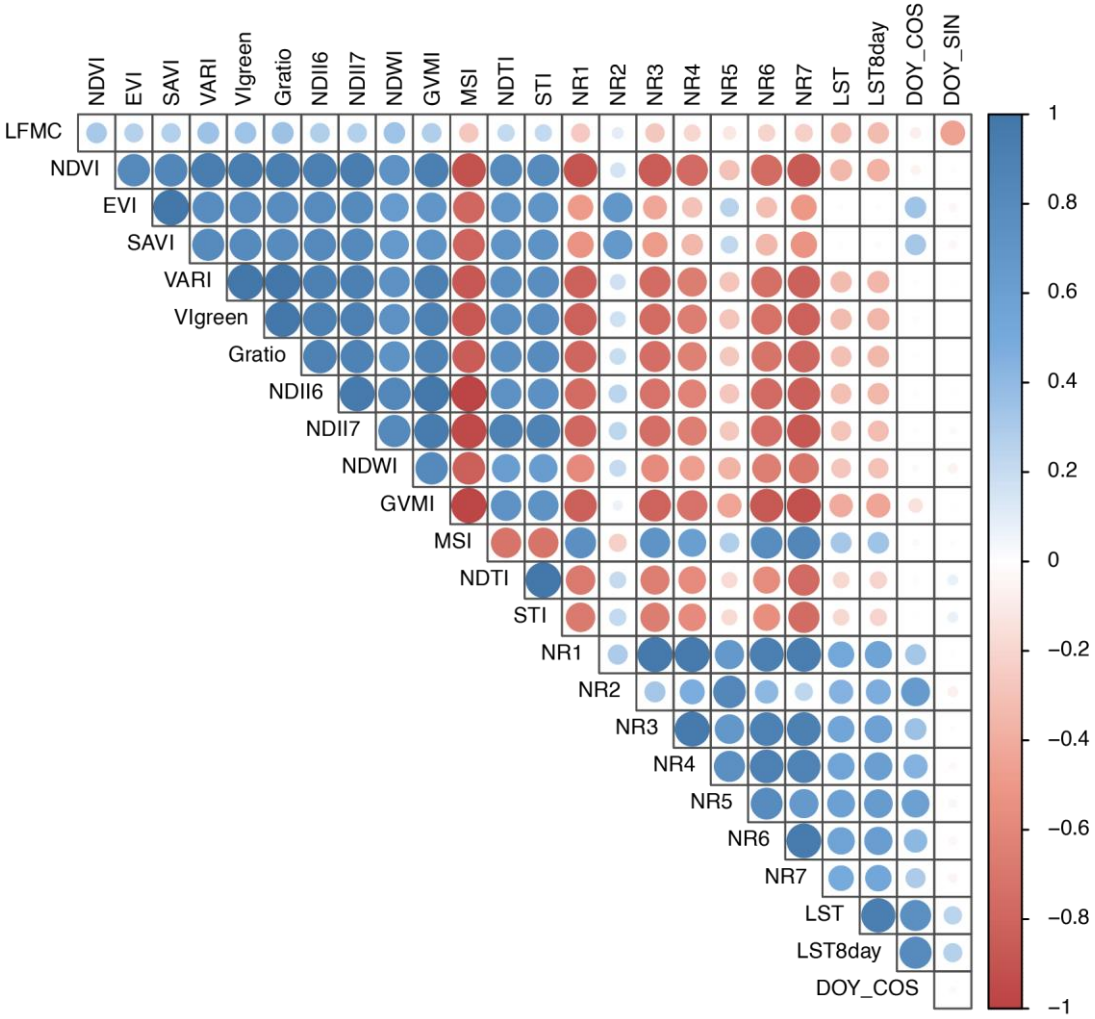

Fig. S3. Correlation matrix between LFMC and predictive variables.

S7. $NDVI_{CV}$ filter

The optimal values for the application of the $NDVI_{CV}$ filter were in the range of 0.3-0.4 (Fig. S4). We additionally examined LFMC predictions made with the calibrated $LFMC_{RF}$ model against observations that were discarded by a $NDVI_{CV}$ threshold value of 0.3 (Fig. S5). The most error estimates were around the mean absolute error of the model (MAE = 15.10%). So, the filter did not discriminate bad quality predictions at all.

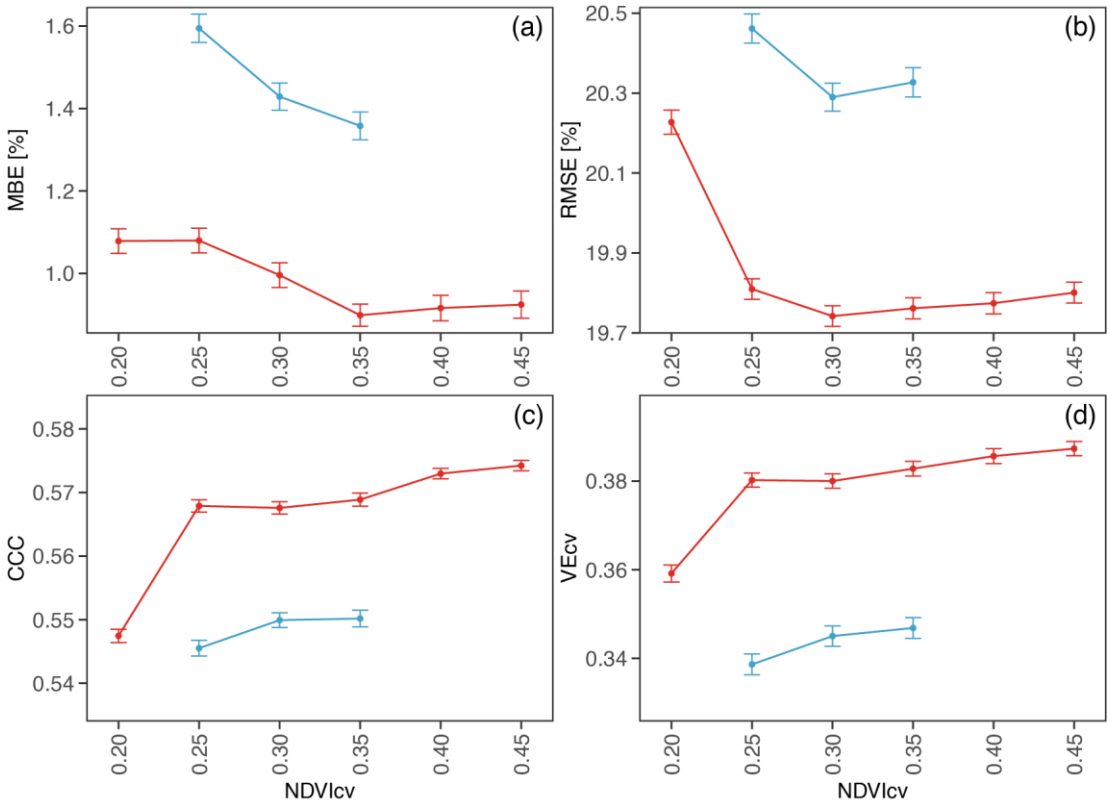

Fig. S4. Performance metrics profiles from the general model performance assessment (MP) alternative with the selected variables and the $NDVI_{CV}$ filter applied to the entire dataset (blue line) and only to the training partition (red line). Dots and vertical segments represent the average value and ±1 standard error obtained from the 100 nested LLOCV repetitions, respectively.

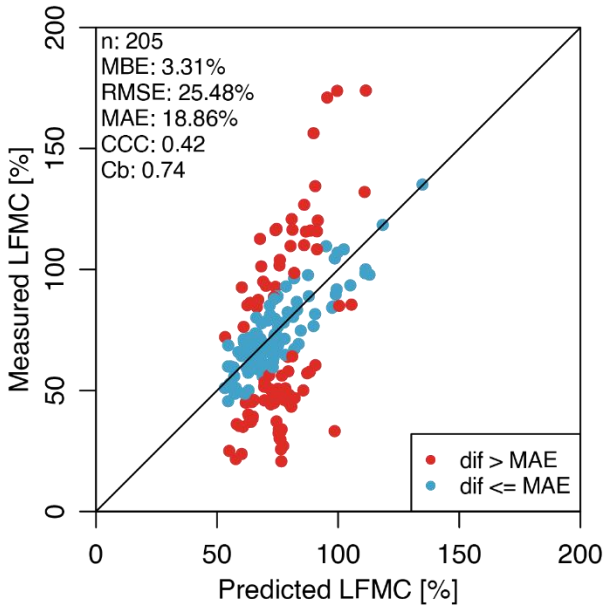

Fig. S5. LFMC field observations versus predictions from the CAL validation theoretically rejected by the 0.3 $NDVI_{CV}$ threshold.

S8. Additional prediction analysis

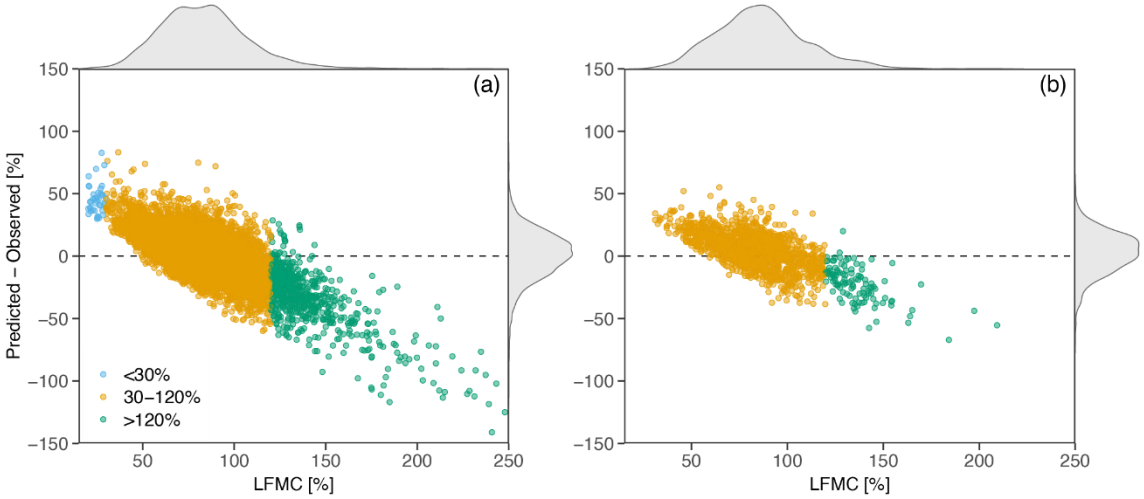

Fig. S6. Residuals between predictions and observations against the LFMC observations and their marginal density distributions for CAL (a) and EXT (b). Point colors highlights LFMC observations below, within and above the critical interval for live fuel flammability.

References

Congalton, R. G., & Green, K. (2019). *Assessing the Accuracy of Remotely Sensed Data*. CRC Press. https://doi.org/10.1201/9780429052729

Meyer, H., Reudenbach, C., Wöllauer, S., & Nauss, T. (2019). Importance of spatial predictor variable selection in machine learning applications – Moving from data reproduction to spatial prediction. *Ecological Modelling*, *411*(March), 108815. https://doi.org/10.1016/j.ecolmodel.2019.108815