Supplementary information for:
Bare soil albedo at high spatial and temporal resolution from Sentinel-2 observations over Europe

Ke Yu[1], Yang Su[1,2,3,*], Philippe Ciais[1], Ronny Lauerwald[2,4], David Makowski[5], Tianqi Shi[1], Shengbiao Wu[6], Petra Sieber[7], Chuanlong Zhou[1], Daniel S. Goll[1]

[1]Laboratoire des Sciences du Climat et de l'Environnement, LSCE/IPSL, CEA-CNRS-UVSQ, Université Paris-Saclay, 91191 Gif-sur-Yvette, France.
[2]UMR ECOSYS, INRAE AgroParisTech, Université Paris-Saclay, 91120 Palaiseau, France
[3]Département d'Informatique, École Normale Supérieure - PSL, 75005 Paris, France
[4]Department Geoscience, Environment & Society-BGEOSYS, Université Libre de Bruxelles, Bruxelles, Belgium
[5]UMR MIA PS, INRAE AgroParisTech, Université Paris-Saclay, 91120 Palaiseau, France
[6]Future Urbanity & Sustainable Environment (FUSE) Lab, Division of Landscape Architecture, Department of Architecture, Faculty of Architecture, The University of Hong Kong, Pokfulam, Hong Kong Special Administrative Region, China
[7]Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland

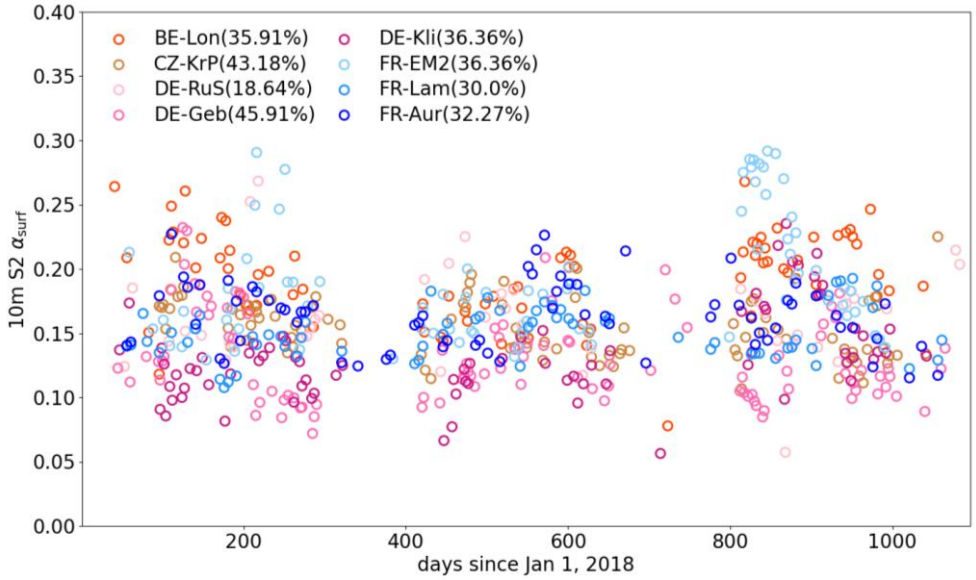
Table of contents

Supplementary Text

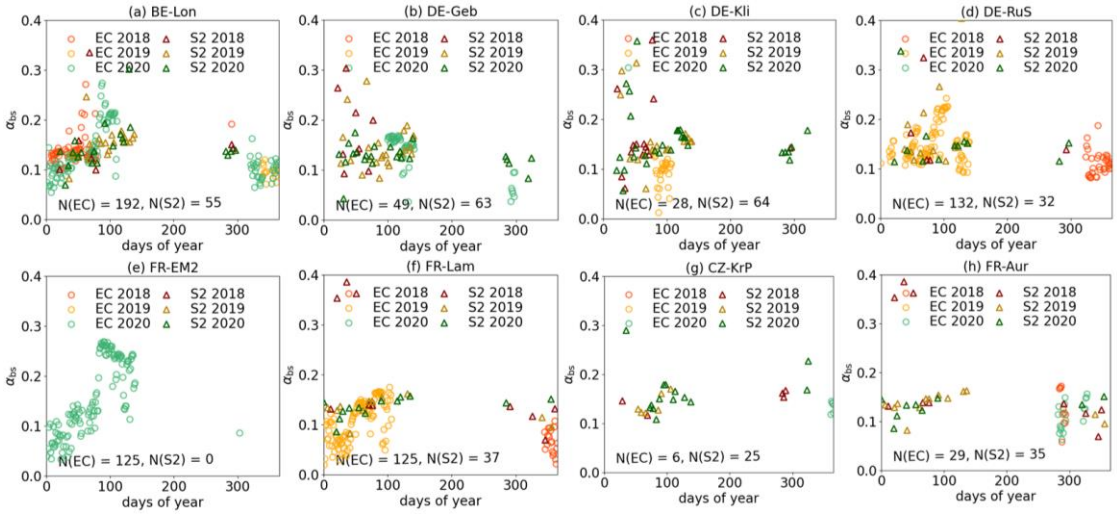
Supplementary Text 1.1 The 10-fold cross-validation in the random forest model

The 10-fold cross-validation divided the training dataset into 10 subsets: 90% for the training model with specified hyperparameters and 10% for validation in each subset. The trained model then predicted the dependent variable using selected features in the 10% validation dataset, and both the dependent variable and the selected features were recorded. During the process, the importance of each feature is calculated based on how much each feature contributes to decreasing the impurity (measured by Gini index) in the tree nodes across all the trees in the forest. Higher scores indicate more important predictors (Strobl et al., 2008). After processing all 10 subsets, the overall hyperparameter-specific coefficient of determination (R^2) and root mean square error (RMSE) were calculated based on the aggregated dependent variable and the selected features from all subsets. The above procedure was repeated for all pairs of hyperparameters and the ones with the highest R^2 were selected to define the final random forest model. Note that the RMSE values for all hyperparameter pairs remain consistently around 0.02.

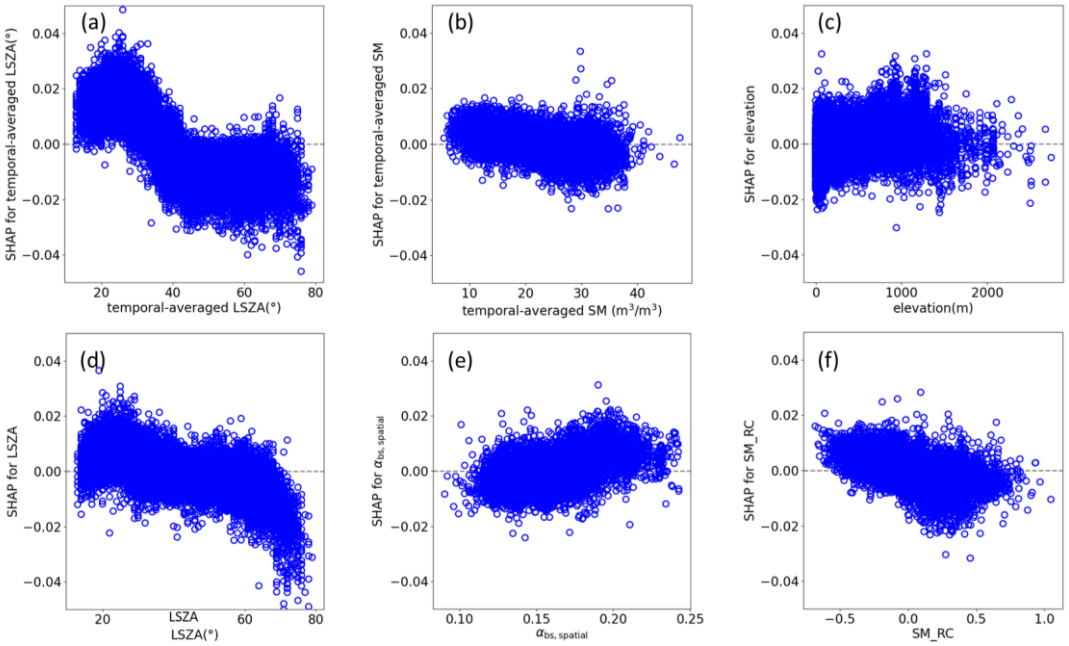
Supplementary Figures



Supplementary Figure 1 The dynamics of 10m 5-day Sentinel-2 surface albedo at 8 ICOS sites from 2018 to 2020. The numbers in legend brackets represent the percentage of available data (%) in three years.

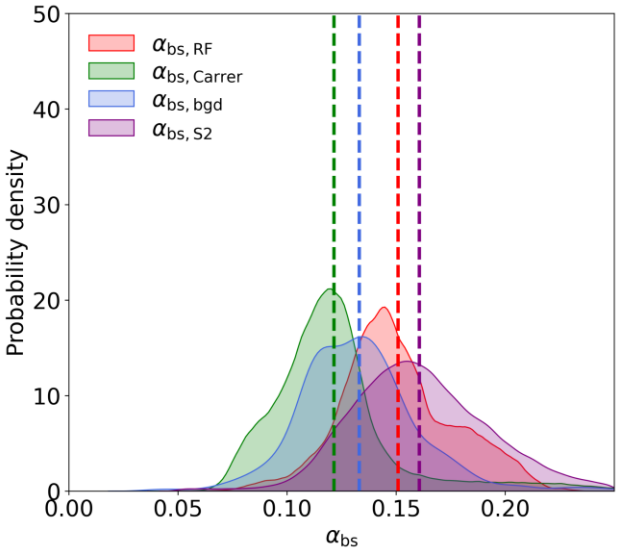


Supplementary Figure 2 The bare soil albedo extracted from Sentinel-2 (S2) surface albedo product ($\alpha_{bs,\ S2}$, triangles) and computed by site eddy covariance (EC) measurements ($\alpha_{bs,\ site}$, circles) during the same bare soil periods detected by field photos instead of satellite vegetation indices at 8 ICOS sites from 2018 to 2020 (Table 3). 'EC' and 'S2' in each plot mean eddy covariance and Sentinel-2, respectively. "N" is the number of dots



67

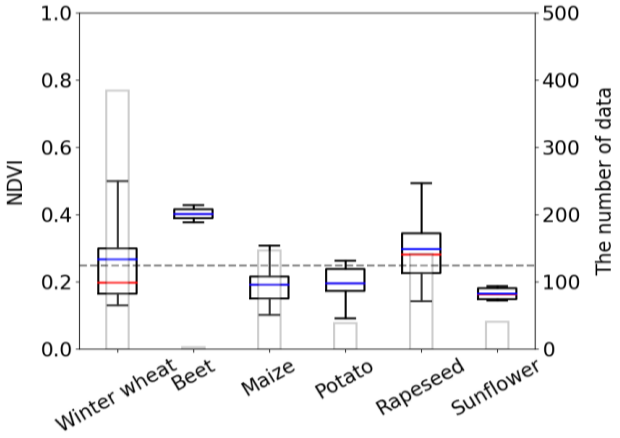
68 **Supplementary Figure 3** The SHapley Additive exPlanations (SHAP) Values of the top 3
69 selected features in the spatial random forest model ((a)-(c)), and in the temporal random forest
70 model ((d)-(f)). 'LSZA', 'SM' and 'SM_RC' represent the local solar zenith angle, soil
71 moisture and relative SM anomaly, respectively. $\alpha_{bs,\ spatial}$ is the spatial bare soil albedo
72 distribution predicted by the spatial random forest model. The blue dots represent the
73 contribution of each feature to bare soil albedo variation.



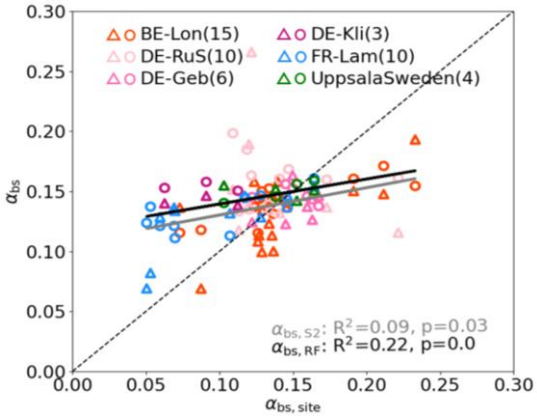
74

75 **Supplementary Figure 4** The probability density of different bare soil albedo (α_{bs}) products
76 within the European cropland regions. The $\alpha_{bs, RF}$ is produced by the trained random forest
77 models from 2018 to 2020, with 300 m spatial resolution and 5-day intervals. $\alpha_{bs, Carrer}$ and $\alpha_{bs, bgd}$ are derived from 8-day 1 km α_{bs} map (Carrer et al., 2014) and a monthly 50 km α_{bs} map
79 (Pinty et al., 2011), respectively. $\alpha_{bs, S2}$ is the α_{bs} datasets trained in random forest models
80 obtained from Sentinel-2 surface albedo observation (Lin et al., 2023) using vegetation indices
81 (see Methods).

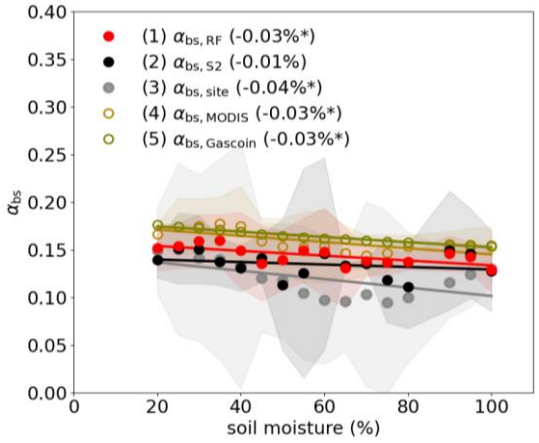
82



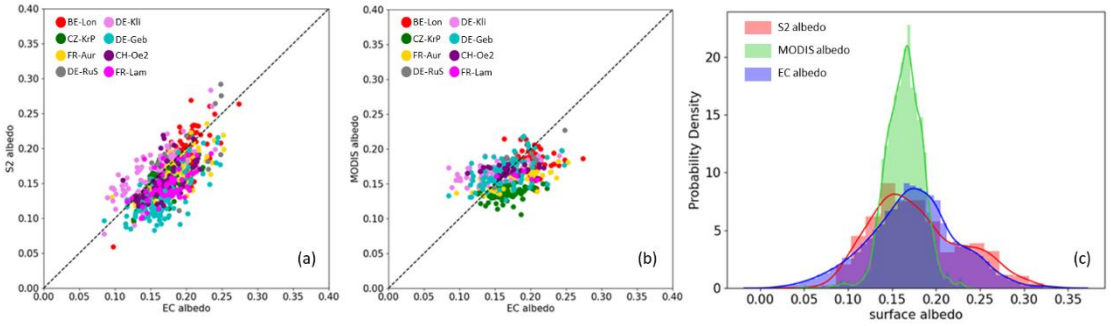
Supplementary Figure 5 The NDVI distribution of crop residues at 6 European cropland sites from 2018 to 2020. The NDVI datasets were derived from 5-day Sentinel-2 reflectance product with a spatial resolution of 300 m on the Google Earth engine platform. The residue periods were identified by daily or weekly site photos. The red and blue lines in each figure represent the median and mean NDVI, respectively. The gray line represents the NDVI threshold of 0.25 used to extract bare soil periods and the corresponding Sentinel-2 bare soil albedo samples.



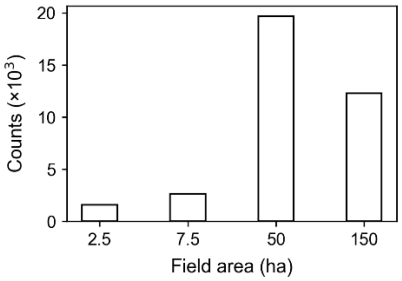
Supplementary Figure 6 The linear relationship between $\alpha_{bs, RF}$ and $\alpha_{bs, site}$ (circles, black line), and between $\alpha_{bs, S2}$ and $\alpha_{bs, site}$ (triangles, gray line) at 4 sites during the same bare soil periods from 2018 to 2020. The bare soil periods are detected by field photos instead of satellite vegetation indices. The $\alpha_{bs, RF}$ is produced by the trained random forest models. The $\alpha_{bs, S2}$ is extracted from the Sentinel-2 surface albedo products, while $\alpha_{bs, site}$ is computed by site eddy covariance measurements (Table 3). Different colors represent different sites. The gray and black lines are the fitting trends between $\alpha_{bs, site}$ and $\alpha_{bs, S2}$, and between $\alpha_{bs, site}$ and $\alpha_{bs, RF}$, respectively. The numbers in the legend brackets represent the sample size. The black dotted line is 1:1 line. R^2 and p are the coefficient of determination and significance level.



Supplementary Figure 7 The linear relationship between $\alpha_{bs, site}$ and soil moisture (SM), $\alpha_{bs, RF}$ and SM, $\alpha_{bs, MODIS}$ and SM, and $\alpha_{bs, Gascoin}$ and SM, respectively (see Table 3). The superficial SM dataset in the x-axis is derived from daily soil sensor measurements which have an SM step of 5% during the same bare soil periods from 2018 to 2020. The dots with areas in different colors show the averages and 2-time standard deviations of all datasets in each SM step. The $\alpha_{bs, RF}$ is produced by the trained random forest models using daily GSSM SM product as one of the predictors (red dots). The $\alpha_{bs, S2}$ is extracted from the Sentinel-2 surface albedo products (black dots), while $\alpha_{bs, site}$ is computed by site eddy covariance measurements (gray dots). The $\alpha_{bs, MODIS}$ is extracted from 500m daily MODIS surface albedo product (MCD43A3) (yellow hollow dots). $\alpha_{bs, Gascoin}$ is computed by the function of SM using the 1km GSSM SM product (brown dots). '*' means the linear relationship (solid lines) passes the significance level of 0.05.



Supplementary Figure 8 The comparison of surface albedo at 8 cropland sites during the same periods between (a) Sentinel-2 (300m, 5-day) and site measurements; (b) MODIS surface albedo product (500m, daily) and site measurements. (c) represent the probability density of the surface albedo from Sentinel-2 data, MODIS product and site measurements.



121

122 **Supplementary Figure 9** The count of field sizes of LUCAS 2018 farms.

References

Carrer, D., Meurey, C., Ceamanos, X., Roujean, J. L., Calvet, J. C., and Liu, S. L.: Dynamic mapping of snow-free vegetation and bare soil albedos at global 1 km scale from 10-year analysis of MODIS satellite products, Remote Sens. Environ., 140, 420-432, doi: 10.1016/j.rse.2013.09.005, 2014.

Lin, X., Wu, S., Chen, B., Lin, Z., Yan, Z., Chen, X., Yin, G., You, D., Wen, J., Liu, Q., Xiao, Q., Liu, Q., and Lafortezza, R.: Estimating 10-m land surface albedo from Sentinel-2 satellite observations using a direct estimation approach with Google Earth Engine, ISPRS J. Photogramm. Remote Sens., 194, 1-20, doi: 10.1016/j.isprsjprs.2022.10.001, 2022.

Pinty, B., Andredakis, I., Clerici, M., Kaminski, T., Taberner, M., Verstraete, M. M., Gobron, N., Plummer, S., and Widlowski, J.-L.: Exploiting the MODIS albedos with the Two-stream Inversion Package (JRC-TIP): 1. Effective leaf area index, vegetation, and soil properties, J. Geophys. Res., 116, D09105, doi: 10.1029/2010JD015372, 2011.

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., and Zeileis, A.: Conditional variable importance for Random Forests, BMC Bioinformatics, 9, 307, doi: 10.1186/1471-2105-9-307, 2008.