Supplementary materials for: "A robust gap-filling approach for ESA CCI soil moisture by integrating satellite observations, model-driven knowledge and machine learning"

Table S1. Summary of the dataset for the preliminary analysis but not the final utilization of the proposed model.

Aims	Variables	Source	Resolution (spatial/temporal)
Model preliminary analysis	EVI	MOD13C1, MYD13C1	0.05°/16 day
	LAI	MCD15A2H	500m/8 day
	Air temperature Solar radiation Wind	Chinese regional ground meteorological dataset	0.1°/3 hourly

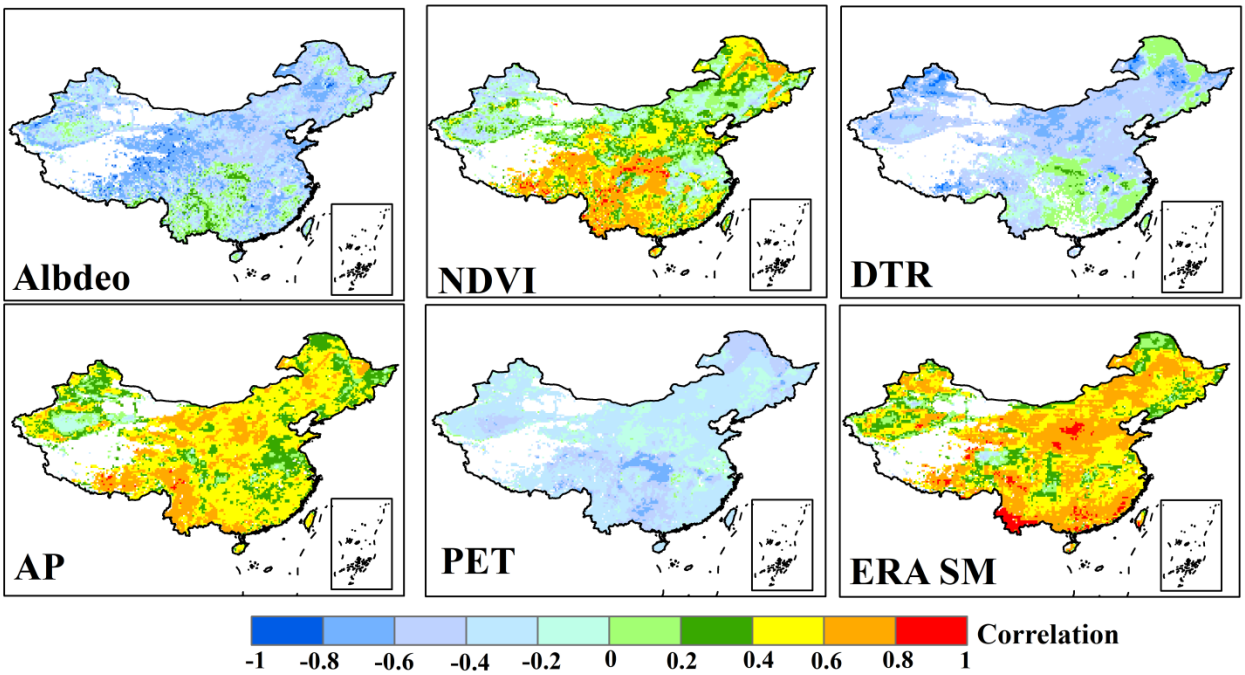

Figure S1. The spatial patterns of correlation between ESA CCI SM and selected variables in 2009

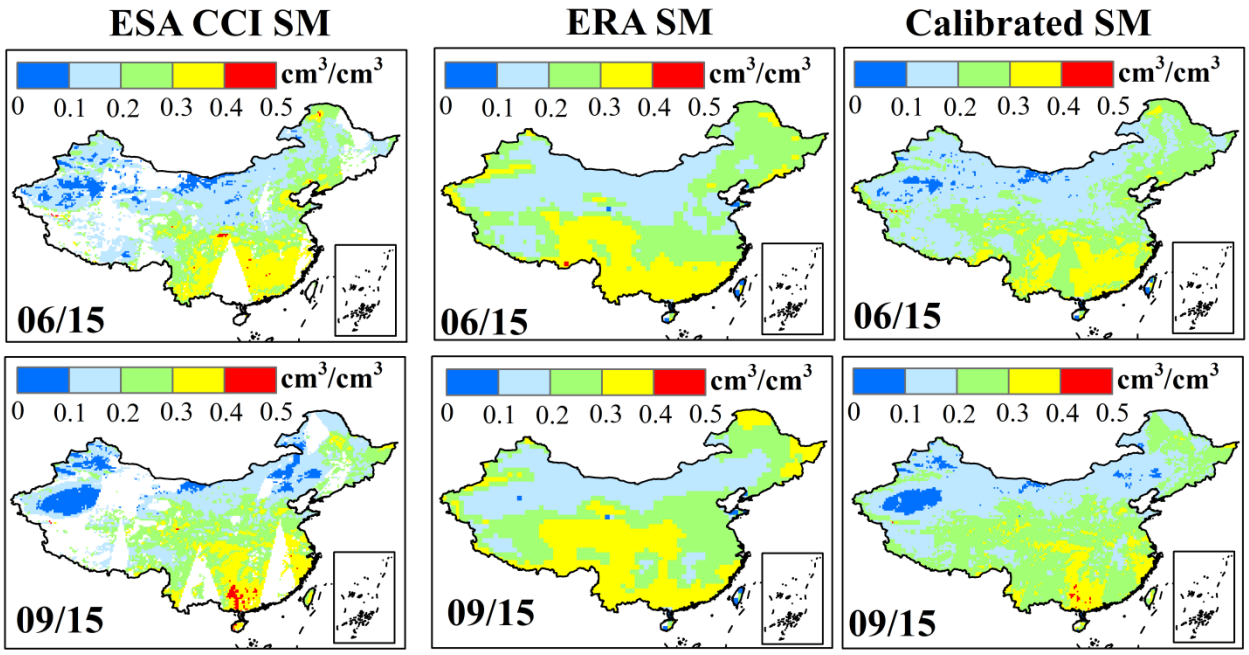

Figure S2. The spatial patterns of ESA CCI SM, ERA5 SM and calibrated ERA SM on the selected days of 2009

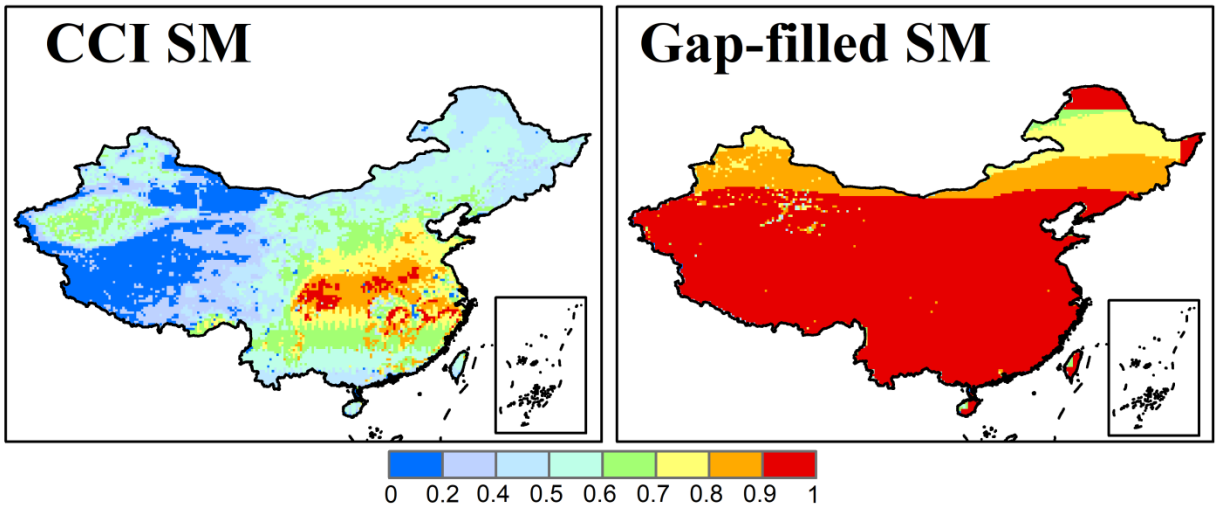

Figure S3. The availability of the original CCI SM and gap-filled SM in 2009

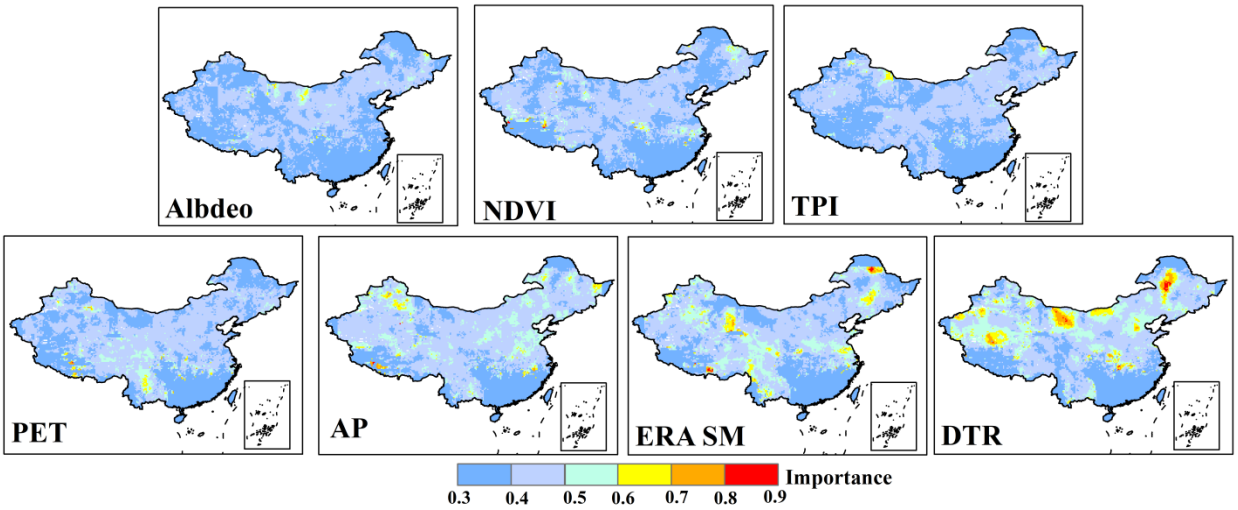

Figure S4. The spatial patterns of the importance of selected variables in 2009

Text S1: The description of DisPATCH model

The DISaggregation based on Physical And Theoretical scale Change (DISPATCH) algorithm is implemented to disaggregate ESA CCI-derived SM. The disaggregation principle beneath this model can be expressed as:

$$SM_H = SM_L + \frac{\delta SM}{\delta SEE} \times (SEE_H - \overline{SEE_H}) \quad (1)$$

where SM_L is low resolution soil moisture (e.g., ECA CCI SM), SM_H is downscaled high resolution soil moisture. SEE_H is the evaporative efficiency retrieved at high resolution scale, and $\overline{SEE_H}$ is the average value within high resolution pixels. $\frac{\delta SM}{\delta SEE}$ is the partial derivative obtained at a low resolution scale. SEE_H is described as

$$SEE_H = \frac{T_{s,max} - T_s}{T_{s,max} - T_{s,min}} \quad (2)$$

with T_s is soil temperature, $T_{s,max}$ and $T_{s,min}$ is soil temperature in dry and wet conditions, respectively. High resolution soil temperature is calculated as

$$T_s = \frac{T_H - f_v T_v}{1 - f_v} \quad (3)$$

where T_H is high resolution land surface temperature (e.g., MODIS), f_v is fractional vegetation cover and T_v denotes vegetation temperature. can be calculated following the studies of Moran et al. (1994).

Text S2: The description of traditional models

Four models are used for comparison analysis, including the Multiple linear regression (MLR), Extreme gradient boost (XGB), Support vector machine (SVM) and Artificial Neural Network (ANN).

1. Multiple linear regression (MLR)

The MLR model can be described as follows:

$$SM = a + \sum x_i \times V_i \quad (4)$$

where SM is reconstructed soil moisture, V is a continuous explanatory variable. The parameter a is intercept value, and x is the regression coefficients.

2. Extreme gradient boost (XGB)

XGB model is an ensemble decision tree model that is implemented based on an advanced gradient boosting framework. A forward fractional algorithm is used in XGB to achieve learning optimization. Specifically, the new regression tree is sequentially generated based on the errors of previous ensemble models, and further trained to literately minimize the cost function. A regular term is added to the cost function for controlling the model complexity, mainly by reducing the model variance.

3. Support vector machine (SVM)

SVM is a robust machine learning algorithm, which is based on an optimization theory. This model is implemented primarily by establishing a set of hyperplanes with maximal margins. The overall SVM can be described as follows:

$$\mathrm{y} = \sum_{i=1}^{M} a_i K(x_i, x) - b \quad (5)$$

where x is the independent vector, and x_i are the trained vectors, M is the number of training data. a_i and b are parameters that can be obtained by maximizing the objective function. K is the kernel function that can simplify the learning process. Here we used the radial based kernel function.

4. Artificial Neural Network (ANN)

The artificial neural network implemented with Levenberg-Marquardt training strategy (Lera and Pinzolas, 2002) is used to conduct SM reconstruction. The activation function used for the hidden layer and output layer is sigmoid purelin, respectively. The output layer is generated with a linear function, which can be described as follows:

$$O = (\sum_{p=1}^{M} i_p \times w_p + b) \times h(x) \quad (6)$$

$$h(x) = \frac{1}{1+e^{-x}} \quad (7)$$

where O is the output of the object hidden layer node, i_p is an input,M is the number of nodes, , w_p is the weight, and b is the bias. $h(x)$ is the sigmoid activation function.

References

Lera, G., Pinzolas, M., 2002. Neighborhood based Levenberg-Marquardt algorithm for neural network training. IEEE Transactions on Neural Networks 13, 1200-1203.

Moran, M.S., Clarke, T.R., Inoue, Y., Vidal, A., 1994. Estimating crop water deficit using the relation between surface-air temperature and spectral vegetation index. Remote Sensing of Environment 49, 246-263.