Supplement of Hydrol. Earth Syst. Sci., 27, 577–598, 2023
https://doi.org/10.5194/hess-27-577-2023-supplement

Hydrology and Earth System Sciences
Open Access
EGU

Supplement of

A robust gap-filling approach for European Space Agency Climate Change Initiative (ESA CCI) soil moisture integrating satellite observations, model-driven knowledge, and spatiotemporal machine learning

Kai Liu et al.

Correspondence to: Shudong Wang (wangsd@aricas.ac.cn)

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	0.05º/16 day
	LAI	MCD15A2H	500m/8 day
	Air temperature	China Meteorological Forcing Dataset	0.1º/3 hourly
	Solar radiation		
	Wind		

Table S2 Summary of the characteristics of in situ sites

ID	Site	Land-use	Elevation	Longitude	Latitude	Soil depth	Projections and references
1	Yucheng	Cropland	23m	116.57E	36.83N	10cm	China Watershed Allied Telemetry Experimental Research (WATER), (Zhang et al., 2021) (Li et al., 2009) (Huang et al., 2016)
2	Daxing	Cropland	20m	116.42E	39.62N	5cm	
3	Miyun	Woodland	350m	117.32E	40.63N	5cm	
4	Guantao	Cropland	30m	115.12E	36.51N	2cm	
5	Arou	Grassland	2995m	100.46E	38.04N	10cm	
6	Maliantan	Grassland	2817m	100.30E	38.55N	5cm	
7	Yingke	Cropland	1519m	100.42E	38.85N	5cm	
8	Guantan	Woodland	2835m	100.25E	38.53N	5cm	
9	AKA	cropland	1008m	80.85E	40.67N	10cm	Chinese Ecosystem Research Network (CERN), (Yu et al., 2006) (Li et al., 2018) (Zhu et al., 2007) (Yao et al., 2018)
10	ALF	Woodland	2455m	101.02E	24.54N	5cm	
11	ASA	cropland	1296m	109.31E	36.85N	10cm	
12	BJF	Woodland	1162m	115.43E	39.97N	5cm	
13	BNF	Woodland	722m	101.02E	21.95N	10cm	
14	CBF	Woodland	512m	127.09E	42.40N	5cm	
15	CLD	Desert	1342m	80.70E	37.01N	10cm	
16	CSA	cropland	21m	120.38E	35.25N	10cm	

17	CWA	cropland	1241m	107.67E	35.25N	10cm	
18	DHF	Woodland	412m	112.53E	23.17N	15cm	
19	ESD	Desert	1301m	110.18E	39.50N	10cm	
20	FKD	Desert	578m	88.00E	44.15N	10cm	
21	FQA	cropland	65m	114.55E	35.02N	10cm	
22	GGF	Woodland	6967m	101.88E	29.60N	10cm	
23	HBG	Grassland	3321m	101.33E	37.66N	5cm	
24	HJA	cropland	305m	108.20E	24.40N	10cm	
25	HLA	cropland	221m	126.63E	47.43N	10cm	
26	HSF	Woodland	102m	112.90E	22.70N	10cm	
27	HTF	Woodland	294m	109.75E	26.83N	10cm	
28	LCA	cropland	52m	114.68E	37.88N	10cm	
29	LSA	cropland	4230m	91.33E	29.66N	5cm	
30	LZD	cropland	1363m	100.12E	39.33N	10cm	
31	MXF	Woodland	2035m	103.90E	31.70N	10cm	
32	NMD	Desert	348m	120.70E	42.92N	10cm	
33	QYA	cropland	48m	115.07E	26.74N	10cm	
34	SNF	Woodland	1611m	110.40E	31.50N	10cm	
35	SPD	cropland	1413m	104.95E	37.45N	10cm	
36	SYA	cropland	35m	123.40E	41.52N	10cm	
37	TYA	cropland	62m	111.50E	28.91N	10cm	
38	YGA	cropland	448m	105.45E	31.27N	10cm	
39	YTA	cropland	44m	116.92E	28.25N	10cm	
40-59	Maqu network	Grassland	~3430m	101.63-102.75E	33.5-34.25N	5cm	Tibetan Plateau observatory of plateau scale soil moisture and soil temperature (Tibet-Obs), (Su et al., 2013) (Wei et al., 2019)

60-716	Agro-meteoroloical stations	Cropland	-84-4200m	75.98-134.28E	18.5-51.72N	10cm	China's agrometeorological observation network, (Meng et al., 2021) (Wang et al., 2016)

Table S3 Optimal parameters regarding seven climate regions

Climate region	n_estimators	max_depth	min_samples_split	max_features
Arid	69	11	8	0.12
Semi-arid	80	18	9	0.16
Aird/semi-wet	47	9	5	0.31
Wet/semi-arid	36	10	3	0.25
Wet	52	15	11	0.16
Moist	62	10	9	0.12
Over-wet	22	8	4	0.27

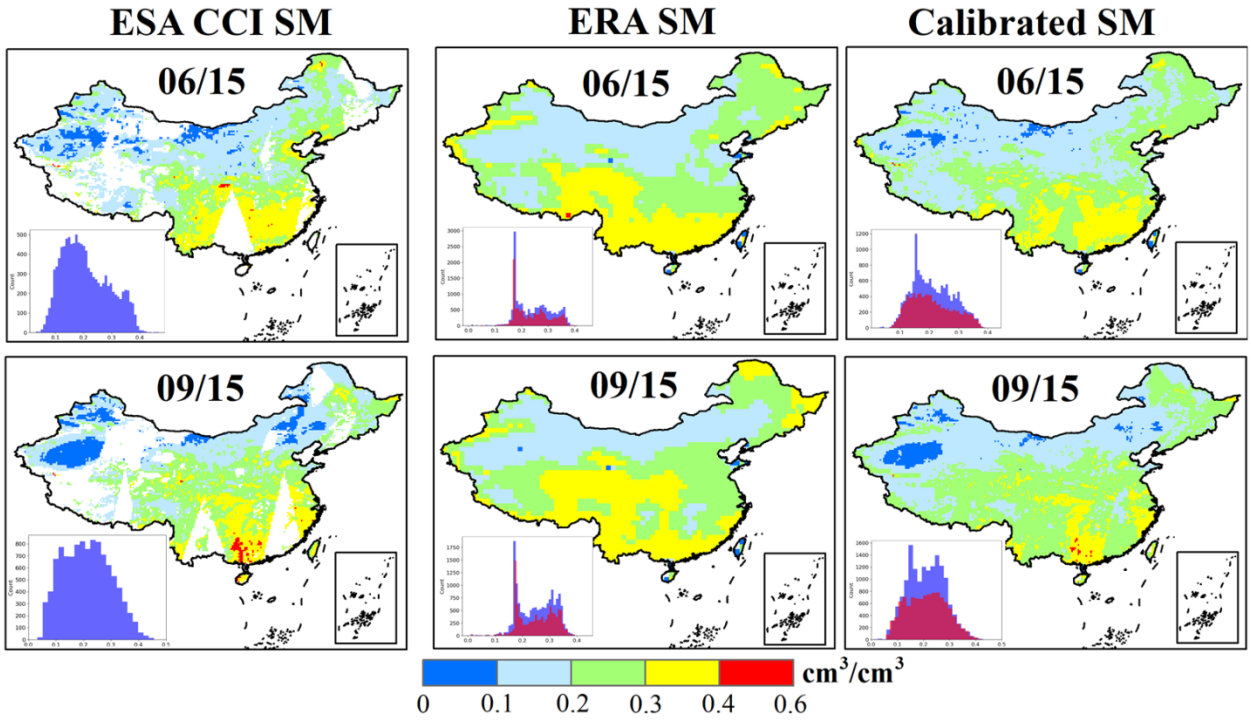

Figure S1. Spatial distributions of ESA CCI SM, ERA5 SM and calibrated ERA SM on the selected days of 2009. The lower-left panel in each sub-figure shows the histogram, and the blue color represents the pixels in which the ESA dataset are available while the red color represents the pixels in which the ERA dataset are available.

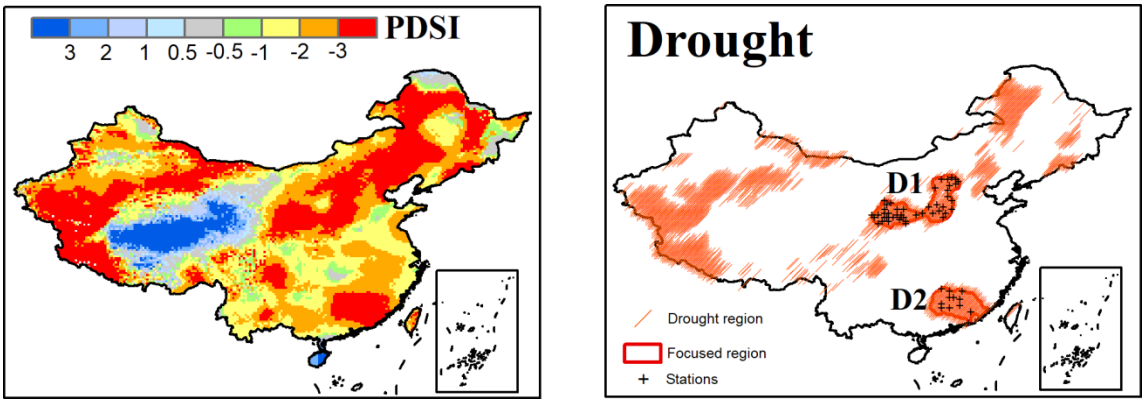

Figure S2. (a) Annual PDSI in 2009. (b) Spatial distribution of drought events, and two severe drought events (D1 and D2) selected for further analysis.

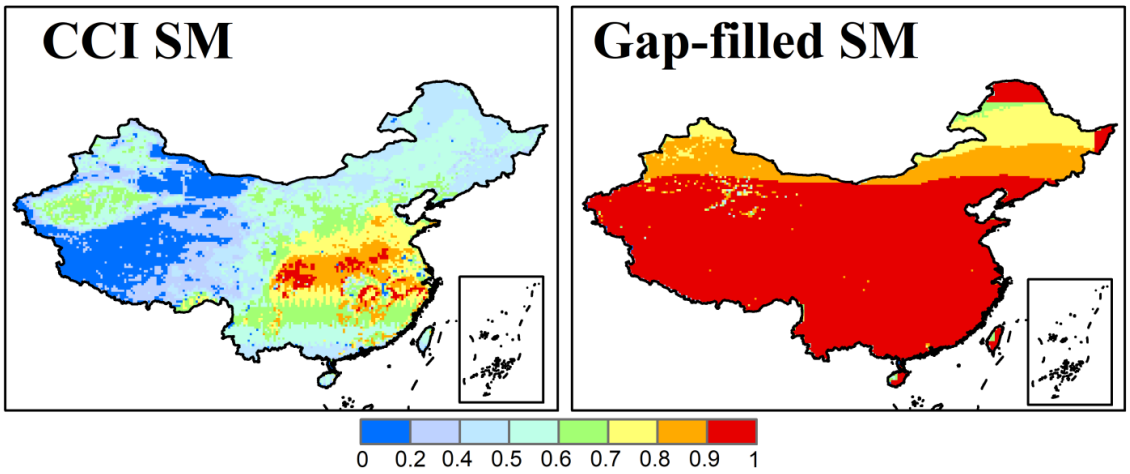

Figure S3. Spatial distribution of availability of the original CCI SM and gap-filled SM in 2009.

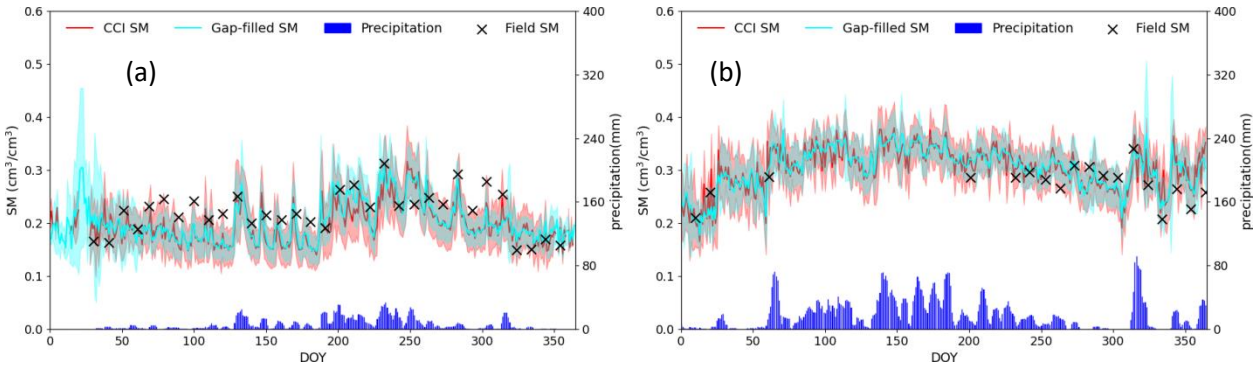

Figure S4. Time series in the (a) region D1 and (b) D2. D1 and D2 are identified in Figure S2.

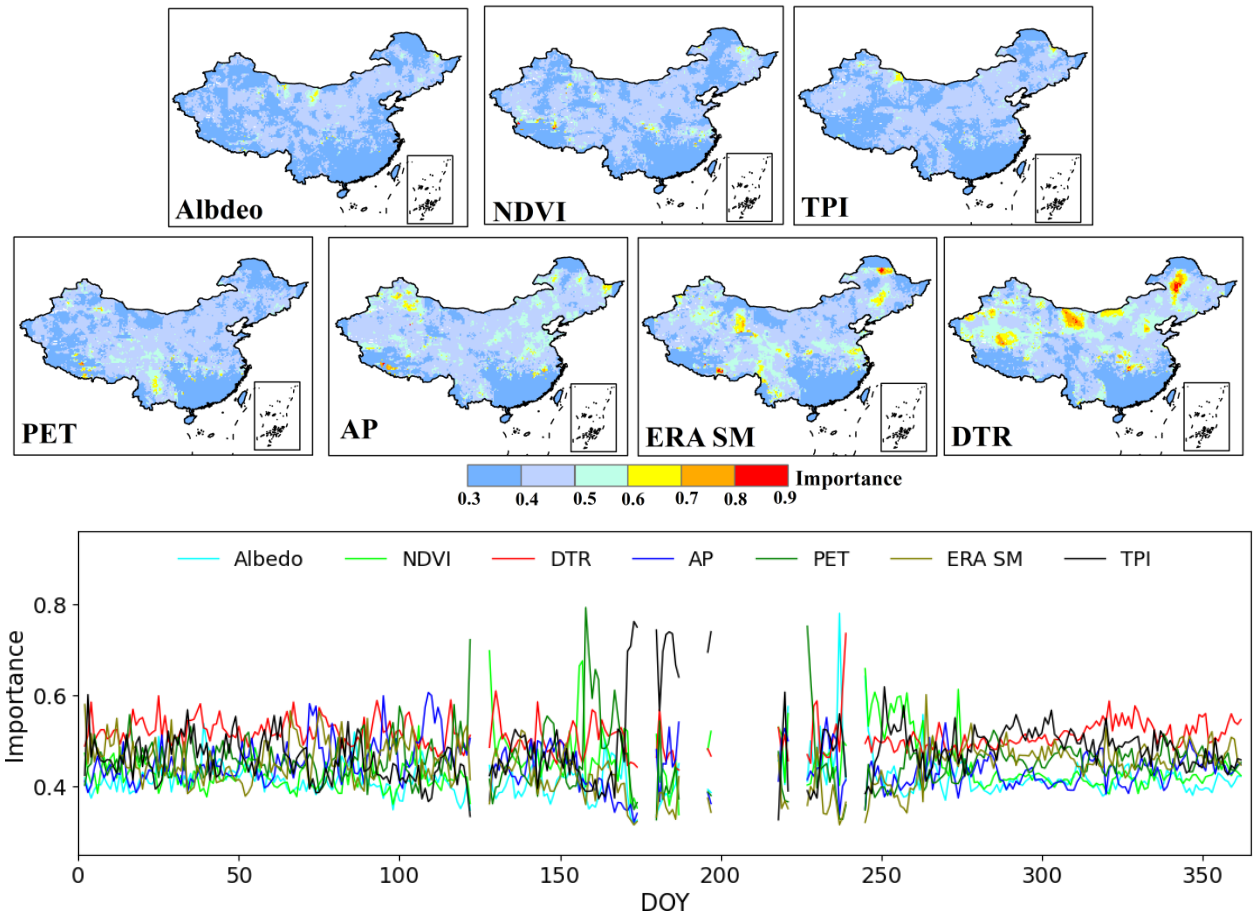

Figure S5. Spatial distributions and time series of the importance score of selected variables in 2009.

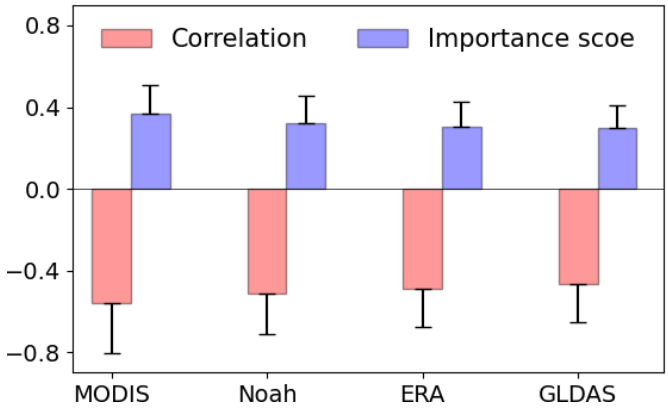

Figure S6. Pearson correlation and importance score of using Noah, ERA, and GLDAS DTR replacing MODIS DTR.

Sect. S1: The regression subset selection approach

The main assumption beneath this regression subset selection approach is that the suppressor variables are associated significantly with each other in regression models, although they may be less correlated with the dependent variables. To be specific, this approach can be conducted with the following steps: (1) using least-squares linear regression to check the potential relationships between SM and explanatory variables; (2) applying a backward stepwise (remove) regression to explore the potential explanatory variables based on the Akaike Information Criterion (AIC); (3) exploiting the best models from all variable combination to identify the important variables impacting SM; and (4) quantifying the relative contributions of each explanatory variable to SM based on the determination coefficient.

Sect. S2: The description of DisPATCH model

As one typical SM disaggregation model, DISPATCH has been extensively applied in current studies (Molero et al., 2016; Song et al., 2021). 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).

Sect. S3: 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

Huang, G., Li, X., Ma, M., Li, H., and Huang, C.: High resolution surface radiation products for studies of regional energy, hydrologic and ecological processes over Heihe river basin, northwest China, Agricultural and Forest Meteorology, 230-231, 67-78, https://doi.org/10.1016/j.agrformet.2016.04.007, 2016.

Lera, G. and Pinzolas, M.: Neighborhood based Levenberg-Marquardt algorithm for neural network training, IEEE Transactions on Neural Networks, 13, 1200-1203, 10.1109/TNN.2002.1031951, 2002.

Li, P., Zhang, L., Yu, G., Liu, C., Ren, X., He, H., Liu, M., Wang, H., Zhu, J., Ge, R., and Zeng, N.: Interactive effects of seasonal drought and nitrogen deposition on carbon fluxes in a subtropical evergreen coniferous forest in the East Asian monsoon region, Agricultural and Forest Meteorology, 263, 90-99, https://doi.org/10.1016/j.agrformet.2018.08.009, 2018.

Li, X., Li, X., Li, Z., Ma, M., Wang, J., Xiao, Q., Liu, Q., Che, T., Chen, E., Yan, G., Hu, Z., Zhang, L., Chu, R., Su, P., Liu, Q., Liu, S., Wang, J., Niu, Z., Chen, Y., Jin, R., Wang, W., Ran, Y., Xin, X., and Ren, H.: Watershed Allied Telemetry Experimental Research, Journal of Geophysical Research: Atmospheres, 114, https://doi.org/10.1029/2008JD011590, 2009.

Meng, X., Mao, K., Meng, F., Shi, J., Zeng, J., Shen, X., Cui, Y., Jiang, L., and Guo, Z.: A fine-resolution soil moisture dataset for China in 2002–2018, Earth Syst. Sci. Data, 13, 3239-3261, 10.5194/essd-13-3239-2021, 2021.

Molero, B., Merlin, O., Malbéteau, Y., Al Bitar, A., Cabot, F., Stefan, V., Kerr, Y., Bacon, S., Cosh, M. H., Bindlish, R., and Jackson, T. J.: SMOS disaggregated soil moisture product at 1km resolution: Processor overview and first validation results, Remote Sensing of Environment, 180, 361-376, https://doi.org/10.1016/j.rse.2016.02.045, 2016.

Moran, M. S., Clarke, T. R., Inoue, Y., and Vidal, A.: Estimating crop water deficit using the relation between surface-air temperature and spectral vegetation index, Remote Sensing of Environment, 49, 246-263, https://doi.org/10.1016/0034-4257(94)90020-5, 1994.

Song, P., Zhang, Y., and Tian, J.: Improving Surface Soil Moisture Estimates in Humid Regions by an Enhanced Remote Sensing Technique, Geophysical Research Letters, 48, e2020GL091459, https://doi.org/10.1029/2020GL091459, 2021.

Su, Z., de Rosnay, P., Wen, J., Wang, L., and Zeng, Y.: Evaluation of ECMWF's soil moisture analyses using observations on the Tibetan Plateau, Journal of Geophysical Research: Atmospheres, 118, 5304-5318, https://doi.org/10.1002/jgrd.50468, 2013.

Wang, S., Mo, X., Liu, S., Lin, Z., and Hu, S.: Validation and trend analysis of ECV soil moisture data on cropland in North China Plain during 1981–2010, International Journal of Applied Earth Observation and Geoinformation, 48, 110-121, https://doi.org/10.1016/j.jag.2015.10.010, 2016.

Wei, Z., Meng, Y., Zhang, W., Peng, J., and Meng, L.: Downscaling SMAP soil moisture estimation with gradient boosting decision tree regression over the Tibetan Plateau, Remote Sensing of Environment, 225, 30-44, https://doi.org/10.1016/j.rse.2019.02.022, 2019.

Yao, Y., Liang, S., Cao, B., Liu, S., Yu, G., Jia, K., Zhang, X., Zhang, Y., Chen, J., and Fisher, J. B.: Satellite Detection of Water Stress Effects on Terrestrial Latent Heat Flux With MODIS Shortwave Infrared Reflectance Data, Journal of Geophysical Research: Atmospheres, 123, 11,410-411,430, https://doi.org/10.1029/2018JD029011, 2018.

Yu, G.-R., Wen, X.-F., Sun, X.-M., Tanner, B. D., Lee, X., and Chen, J.-Y.: Overview of ChinaFLUX and evaluation of its eddy covariance measurement, Agricultural and Forest Meteorology, 137, 125-137, https://doi.org/10.1016/j.agrformet.2006.02.011, 2006.

Zhang, C., Long, D., Zhang, Y., Anderson, M. C., Kustas, W. P., and Yang, Y.: A decadal (2008–2017) daily evapotranspiration data set of 1 km spatial resolution and spatial completeness across the North China Plain using TSEB and data fusion, Remote Sensing of Environment, 262, 112519, https://doi.org/10.1016/j.rse.2021.112519, 2021.

Zhu, L., Sun, O. J., Sang, W., Li, Z., and Ma, K.: Predicting the spatial distribution of an invasive plant species (Eupatorium adenophorum) in China, Landscape Ecology, 22, 1143-1154, 10.1007/s10980-007-9096-4, 2007.