A robust gap-filling approach for ESA CCI soil moisture ~~by~~ integrating satellite observations, model-driven knowledge, —and spatiotemporal machine learning

Kai Liu[1], Xueke Li[2], Shudong Wang[1,3], Hongyan Zhang[1]

[1]Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China.
[2]Institute at Brown for Environment and Society, Brown University, Providence, RI, 02912, USA
[3]Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters (CIC-FEMD), Nanjing University of Information Science & Technology, Nanjing 210044, China

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

Abstract. ~~Soil moisture (SM) is a critical component of the water cycle and a key ecological process connecting the soil-vegetation-atmosphere system.~~ Spatiotemporally continuous soil moisture (SM) data are~~is~~ increasingly ~~demanded~~ in demand for ecological and hydrological research ~~fields~~. Satellite remote sensing has ~~opened opportunities~~potential for mapping SM~~.~~, but~~Nevertheless,~~ the continuity of satellite-derived SM ~~imagery~~ is hampered by data gaps, resulting from inadequate satellite coverage and radio-~~-~~frequency interference. ~~In light of this~~Therefore, we propose a new gap-filling approach to reconstruct daily SM time series using the European Space Agency's Climate Change Initiative ~~(ESA CCI)~~. The developed approach integrates satellite observations, model-driven knowledge, and a machine learning algorithm that leverages both spatial and temporal domains. Taking SM in China as an example, ~~we show high accuracy of~~ the reconstructed SM showed high accuracy when validated ~~with~~ against multiple sets of in situ measurements, with ~~a~~ root mean square error (RMSE) and mean absolute error (MAE) of 0.09–~~-~~0.14 and 0.07–~~-~~0.13 cm^3/cm^3, respectively. Further evaluation with a 10-~~-~~fold cross validation ~~reveals~~ revealed ~~a~~ median value of the coefficient of determination (R^2), RMSE, and MAE of 0.56, 0.025 cm^3/cm^3 and 0.019 cm^3/cm^3, respectively. The reconstructive performance ~~is~~ was noticeably reduced both when excluding ~~an~~ one explanatory variable ~~while~~ and keeping the ~~rest~~ other variables ~~remains~~ unchanged, ~~as well as~~and when removing the spatiotemporal domain strategy ~~and~~ or the residual calibration procedure~~, respectively~~. ~~Compared to that~~In comparison with gap-filled SM data based on ~~using~~ a satellite-derived diurnal temperature range (DTR), the gap-filled SM data ~~reconstructed SMs using~~from bias-corrected model-derived DTRs exhibited ~~acceptable~~ relatively lower accuracy ~~accuracies and~~ but higher spatial coverage. ~~Applying~~ Application of our gap-filling approach to long-term SM data-sets (2005—~~-~~2015) produced~~, we show~~ a promising result (~~with a~~ R^2 ~~of~~ = 0.72). A more accurate trend ~~is~~ was achieved relative to that of the original CCI SM when assessed with in situ measurements (i.e., 0.49 versus 0.28, respectively, in terms of R^2). Our findings indicate the feasibility of integrating satellite observations, model-driven knowledge, and spatiotemporal machine learning ~~for~~ to fill~~ing~~ gaps in short- and long-term SM time series ~~over short and long time scales~~, ~~therefore~~ thereby providing a potential avenue for applications to similar studies.

1. Introduction

As an essential component of land–~~-~~atmosphere interactions, soil moisture (SM) substantially impacts the energy, water, and carbon cycles. It plays ~~an~~ important roles in hydrological, environmental, and agricultural applications such as evapotranspiration (ET) estimation (Detto et al., 2006), drought assessment (Wang et al., 2011), and flood forecasting (Wanders et al., 2014). SM has been declared by the Global Climate Observing System (GCOS) and United Nations Framework Convention on Climate Change (UNFCCC) as one of the 50 vital variables in terrestrial domains (Gcos, 2010). Availability of s~~S~~patially and temporally continuous daily all-weather SM data could facilitate improved~~s~~ understanding of ecological and hydrological processes~~,~~; therefore, provision of ~~and~~ a reliable SM dataset is urgently demanded.

~~There are a variety of ways of~~Various methods are available for collecting SM data. In situ measurements can capture the temporal variability of SM at the station scale, and many~~.~~ ~~Plenty of in-situ monitoring~~ networks designed for such in situ observations have been installed regionally, nationally, and globally, e.g., the crop growth and farmland SM database in China, the North American Soil Moisture Database in North America, and the International Soil Moisture Network (ISMN) (Schaake et al., 2004; Dorigo et al., 2011). Nevertheless, owing to the limited number of ground stations, ~~it is challenging to~~ obtaining spatially continuous SM measurements across large-scale regions remains a challenge. In addition to ground-based observations, SM can be simulated ~~with~~ using ~~various~~ numerical models. The Global Land Data Assimilation System (GLDAS) and European Centre for Medium-Range Weather Forecasts (ECMWF) fifth-generation global atmospheric reanalysis (ERA5) can model the soil moisture values that have sufficient spatial coverage (Chen et al., 2013; Reichle et al., 2011). However, ~~these~~ such model ~~simulated~~ simulations ~~dataset~~ tend~~s~~ to be sensitive to ~~the~~ uncertainties ~~relating~~ related to model structure, forcing, and parameterization ~~(Prihodko et al., 2008)~~ (Prihodko et al., 2008; Dorigo et al., 2017).

Satellite observation ~~has been~~is considered ~~as one~~a powerful technique for retrieving surface SM data, especially ~~accompanying~~ given ~~the increasing~~ improvements ~~of~~ in sensor technology ~~(Crow et al., 2012)~~. Some SM-dedicated satellites, e.g., the Advanced Microwave Scanning Radiometer-Earth Observation System (AMSR-E), and Advanced Scatterometer (ASCAT) have used the higher C-band and X-band microwave frequencies to collect SM signals ~~(Paloscia et al., 2001)~~. ~~However, these satellite systems are less~~Despite the sensitivity of satellite-derived SM data ~~subject~~ to atmospheric variability and vegetation coverage~~.~~, ~~Apart from this, some observation sensors are installed with~~satellites operating with ~~the~~ lower L-band radiometers, such as the Soil Moisture and Ocean Salinity (SMOS) (Kerr et al., 2001) and Soil Moisture Active and Passive (SMAP) ~~(Entekhabi et al., 2010).~~ (Entekhabi et al., 2010), ~~These observation systems~~ have exhibited great potential ~~in~~ for collecting SM data ~~sources due to~~because of the strong capacity of wavelengths in the L-band ~~in~~ frequency range to penetrate~~ing~~ vegetation. A case worth noting is that the Climate Change Initiative of the European Space Agency (ESA CCI) has generated one set of global SM dataset (Gruber et al., 2019; Dorigo et al., 2017). This CCI SM product blends a series of SM products from active passive microwave satellite sensors, enabling it one complete and consistent observational SM record. ~~Despite some uncertainties, earlier~~Previous studies have revealed ~~good~~ reasonable

~~accordance~~ correlation between the CCI SM dataset and ~~the~~ in situ measurements obtained over different regions (Dorigo et al., 2015).

The gap issues that remain in current satellite-based SM products relate to a various factors such as radio-frequency interference and orbital changes of the satellite sensors. ~~Although the active and passive microwave satellite sensors can depict soil moisture characteristics across large scales, the gap issues still exist in these satellite-based SM products. This is related to a variety of factors, such as the radio-frequency interference and orbit changes of satellite sensors.~~ Considerable effort~~s~~ ~~have~~ has been dedicated to filling ~~the~~ missing values in ~~the~~ satellite-derived SM datasets. Traditional interpolation approaches that are applied to fill gaps ~~relying~~ rely on the spatial or temporal patterns of the target variable, such as ~~the~~ inverse distance weighting and cokriging (Yao et al., 2013; Ford and Quiring, 2014). ~~Some o~~Other studies have focused on the use of statistical methods that mainly depend on the statistical and physical relationships between target variables and explanatory variables ~~(Leng et al., 2017)~~ (Leng et al., 2017; Llamas et al., 2020; Meng et al., 2021). Only recently m~~M~~achine learning strategies have been ~~recently~~ introduced to the problem of ~~in~~ gap-filling in relation to ~~the~~ satellite-derived dataset (Zhang et al., 2021b; Zhang et al., 2021a). ~~These~~ Such methods have ~~a~~ strong capacity ~~in~~ for depicting complex relationships of target variables and explanatory variables. For instance, Elsaadani et al. (2021) assessed the spatiotemporal deep learning method for filling the gaps in soil moisture observations, (Li et al., 2022c; Li et al., 2021b) further improved satellite soil moisture prediction using deep learning model. ~~Compared to~~ In comparison with statistical-based models,~~ the~~ machine learning models ~~may~~ might be more flexible and robust, especially ~~regarding~~ with regard to ~~the~~ complex scenes and extended coverage (Reichstein et al., 2019).

Most ~~current~~ SM gap-filling studies ~~typically~~ rely on explanatory variables that are required in describing SM dynamics. ~~The common explanatory variables include~~ In addition to satellite-derived vegetation index (e.g., normalized difference vegetation index (NDVI) and enhanced vegetation index (EVI)), surface albedo, and land surface temperature (LST), ~~as well a variety of~~various climatic and geographical factors have been employed in ~~these~~ such studies (Almendra-Martín et al., 2021; Cui et al., 2019; Jing et al., 2018). Nevertheless, although appropriate for use in certain regions, most of ~~these~~ those variables are less suitable for use in heterogeneous regions and extended coverage~~, although they are suitable for regional areas~~. For example, previous studies (Song et al., 2021; Liu et al., 2020b) ~~illustrated~~ that ~~studies focusing~~focused on NDVI and LST ~~tend~~ tended to achieve better performance in delineating SM in arid and semi-arid regions, but ~~produce~~ produced unsatisfactory performance in humid areas. Moreover, ~~these~~ satellite-derived variables (e.g., optical and thermal infrared parameters) are likely to be impacted by cloud conditions. Accordingly, researchers have attempted to explore effective information for promoting model establishment and application. Some studies ~~use~~ used the feature transform approaches to extract distinct signals for driving models. Principal component analysis (PCA) and wavelet decomposition have been employed to reconstruct SM and other satellite-based parameters (Uebbing et al., 2017; Almendra-Martín et al., 2021). Despite ~~pretty good~~ reasonable model performance achieved in ~~the~~ humid and semi-arid region (Zhang et al., 2016; Almendra-Martín et al., 2021), some studies found ~~that there is~~ no substantial improvement in model performance in areas of cropland ~~of~~ in semi-humid ~~region~~ regions when using the PCA (Wang et al., 2020). ~~Some other~~Other studies ~~have~~ focused on the distinct dataset source for gap-filling models. Soil moisture from GLDAS, ERA5, China Meteorological Administration Land Data Assimilation System (CLDAS) and Fengyun Microwave Radiation Imager is considered (Long et al., 2019; Cui et al., 2020). The gap-filling models integrating these unique dataset sources are able to describe SM dynamics, but uncertainties ~~are still observed in~~ ~~the~~ remain in relation to humid regions and areas subject to the freezing-thaw ~~areas~~ process (Song et al., 2021; Cui et al., 2019). Overall, progress regarding the availability of explanatory variables ~~in contributing SM reconstructing models~~ for use in models for reconstruction of SM is inadequately ~~explored, which~~and this is especially critical for machine learning gap-filling models that are sensitive to the structure of the input sequences (Mao et al., 2019).

Although earlier studies ~~have~~ focused on completing ~~the~~ SM dataset, most ~~of them~~ partially ~~aim at the~~addressed a specific case of satellite to observations but ~~less~~ failed to consider ~~the~~ large continental region. Almendra-Martín et al. (2021) and Liu et al. (2020b) applied reconstruction algorithms to the CCI SM product in regional Europe and Oklahoma, USA, respectively, and Cui et al. (2019) continuously promote this approach in the Tibetan Plateau. ~~These~~ Such models rely on machine learning algorithms and a variety of satellite-based variables. Furthermore, ~~studies aimed at~~research on the challenging case of SM time series ~~SM dataset~~ at the daily scale ~~are insufficiently implemented~~ (Zhang et al., 2021b; Long et al., 2019), which is fundamental to ~~explore the~~ the exploration of SM dynamics, and ~~quantify its impacts and~~ the quantification of the associated impact on the contribution ~~on~~ to climate change and water cycle is limited (Bessenbacher et al., 2022a).

~~This study proposes~~Here, we propose a robust gap-filling methodology for ~~reconstructing~~ reconstruction of a spatially continuous daily ESA CCI SM dataset, primarily based on satellite observations, model-driven knowledge and one spatiotemporal random forest algorithm. Our model ~~is applied~~ was tested by application to continental China, which has ~~sufficient landscape~~ suitable variability in terms of landscape and climatic conditions. ~~To be specific~~ Specifically, the feasibility and merit of the developed model ~~are~~ were demonstrated by the following: 1) ~~evaluating~~ evaluation of the gap-filled results ~~with the in-situ~~ using in situ measurements and ~~the~~ holdout ~~cross-validation~~ cross validation, and ~~comparing~~comparison against those of other models~~;~~ , and 2) ~~discussing the~~examination of model uncertainty in terms of the filtered explanatory variables, and extending the proposed model to one long-term period.

2. Study region ~~and material~~

~~2.1 Study region~~

China is located from 3°51′N to 53°33′N and from 73°33′E to 135°05′E, covering an area of approximately 9.6×10^6 km^6 (Fig. 1). A variety of terrain types are presented across China, including the plain, basin, plateau, mountain and hill. These diverse terrains inevitably result in noticeable spatial differences in precipitation and temperature, accompanying the elevation decreasing from west to east. Seven climate zones can be identified in China, including arid, semi-arid, arid/semi-wet, wet/semi-arid, wet, moist, and over-wet climates. The identification of this zoning system is based on a China's humidity index map produced by the National Earth System Science Data Center, National Science & Technology Infrastructure of China (http://www.geodata.cn).

~~In addition to the whole regions of China, we also chose two local regions for model uncertainty analysis (Fig. 1). One region is focused on northern China (NC) which is mostly occupied by arid and semi-arid areas, while the other region is focused on southern China (SC) that is occupied by wet areas.~~

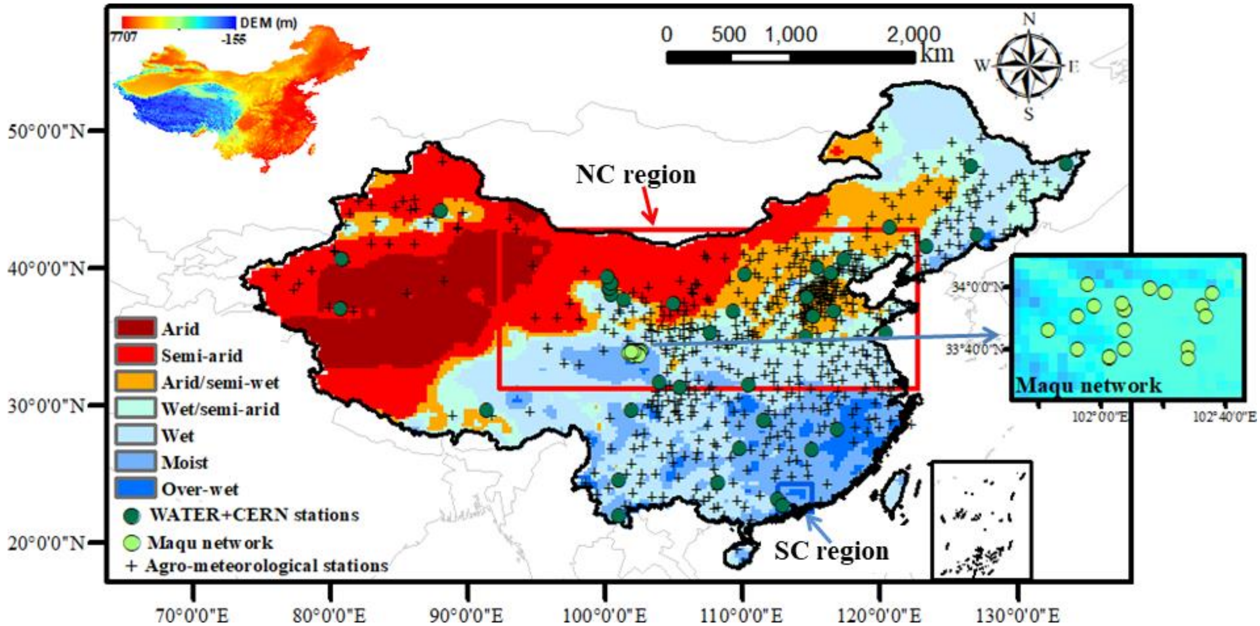


Figure 1: The study region and the selected in situ soil moisture sites. The figure in the upper-left corner ~~delineates the DEM information~~ shows the Digital Elevation Model (DEM) information. The detailed distribution of dense in situ measurements in the Maqu network is shown in the figure on the far right. Two regional areas for uncertainty analysis (i.e., northern China (NC) and southern China (SC)) are ~~delineated with~~ bordered by the ~~rectangle~~ rectangles.

~~2.1 Material~~3. Materials and methods

The object of this study was to reconstruct CCI SM data gaps to produce spatially continuous data records. The basic principle of the proposed gap-filling approach is to efficiently determine the correlation between SM records and the corresponding explanatory variables, which can be expressed as follows:

$$SM = f(V_1, V_2, V_3 \cdots\cdots V_k) + \varepsilon, \quad (1)$$

$$V_i \in R^{N,T}, \quad (2)$$

where SM is the soil moisture, V_i is the corresponding explanatory vectors, and k is the number of the input variables. V_i can be a vector, and the sample number is determined the spatial domain (N) and temporal domain (T). f is one function that can be either linear or nonlinear. ε represents the model residual. In machine learning ensemble, f represents a black box model that does not have one specific form.

The proposed methodology involves three core steps: (i) using a regression subset selection approach and a variable correction procedure to filter explanatory variables from the satellite observations and model-driven knowledge, and to correct the systematic variable bias between them; (ii) training a machine learning algorithm to determine the SM-explanatory variables correlation based on the selected optimal parameters and the available pixels identified with a spatiotemporal window search strategy, and then applying the established correlation to retrieve the unavailable SM pixels; and (iii) conducting geographically weighted regression and Gaussian filtering to calibrate the model-derived residuals. Figure 2 shows a schematic of the overall procedure including the dataset processing, model implementation and model analysis.

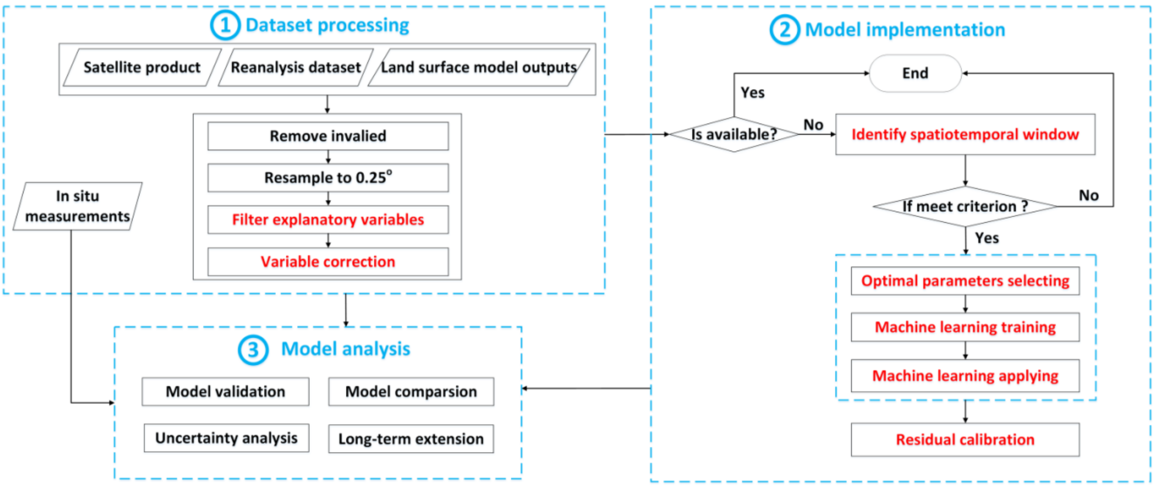


Figure 2: The schematic of overall procedure. The red text denotes the core procedures conducted in the proposed model, which will be described in the following sections.

3.1 Dataset processing

The dataset used ~~mainly ~~includes~~, i)~~ satellite product, reanalysis dataset, ~~and ~~land surface model ~~products for the model establishment, ii)~~outputs and in situ measurements~~ for model validation, and iii) reanalysis dataset and land surface model dataset for model performance analysis. (~~Table 1 ~~(~~and Table S1). ~~provides the information of the used dataset. ~~Details about these datasets are described in the following sections.

Table 1. Summary of the dataset used for the proposed model. Other dataset for the preliminary analysis but not the final utilization of the model is exhibited in supplementary Table S1.

~~Aims~~	~~Variables~~	~~Source~~	~~Resolution (spatial/temporal)~~
	~~Soil moisture~~	~~ESA CCI~~	~~0.25°/daily~~
	~~Surface albedo~~	~~MCD43C3~~	~~0.05°/16-day~~
	~~NDVI~~	~~MOD13C1, MYD13C1~~	~~0.05°/16-day~~
	~~Land surface temperature (LST)~~	~~MYD11C1~~	~~1km/instantaneous~~
~~Model establishment~~	~~Precipitation~~	~~Chinese regional ground meteorological dataset~~	~~0.1°/3-hourly~~

	~~Potential evapotranspiration (PET)~~	~~GLEAM~~	~~0.25°/daily~~
	~~Soil moisture~~	~~ERA5~~	~~0.25°/hourly~~
	~~Land cover classification~~	~~MCD12Q1~~	~~500/annual~~
	~~Digital Elevation Model (DEM)~~	~~SRTM~~	~~90m~~
~~Model analysis~~	~~Surface temperature~~	~~Noah simulations from previous work~~	~~1km/3 hourly~~
	~~Surface temperature~~	~~ERA5~~	~~0.25°/hourly~~
	~~Surface temperature~~	~~GLDAS~~	~~0.25°/3-hourly~~
	~~Soil moisture~~	~~GLDAS~~	~~0.25°/3-hourly~~
	~~Soil moisture~~	~~GLEAM~~	~~0.25°/daily~~
ID	Variables	Source	Resolution (spatial/temporal)
1	Soil moisture	ESA CCI	0.25°/daily
2	Surface albedo	MCD43C3	0.05°/16 day
3	NDVI	MOD13C1, MYD13C1	0.05°/16 day
4	Land surface temperature (LST)	MYD11C1	1km/instantaneous
5	Precipitation	China Meteorological Forcing Dataset	0.1°/3 hourly
6	Potential evapotranspiration (PET)	GLEAM	0.25°/daily
7	Soil moisture	ERA5	0.25°/hourly
8	Land cover classification	MCD12Q1	500/annual
9	Digital Elevation Model (DEM)	SRTM	90m
10	Surface temperature	Noah simulations from previous work	1km/3 hourly
11	Surface temperature	ERA5	0.25°/hourly
12	Surface temperature	GLDAS	0.25°/3-hourly
13	Soil moisture	GLDAS	0.25°/3-hourly
14	Soil moisture	GLEAM	0.25°/daily
15	in situ soil moisture	China Watershed Allied Telemetry Experimental Research (WATER)	daily
16	in situ soil moisture	Chinese Ecosystem Research Network (CERN)	5 day
17	in situ soil moisture	Tibetan Plateau observatory of	daily

		plateau scale soil moisture and soil temperature (Tibet-Obs)	
18	in situ soil moisture	China's agrometeorological observation network	10 daily

~~2.1.1 Satellite dataset~~3.1.1 Satellite product

The ESA CCI SM dataset is provided by the Climate Change Initiative program of the European Space Agency. This product is primarily composed of three types of daily dataset sources, i.e., active, passive, and active-passive combined microwave products (Dorigo et al., 2017). Despite the wide spatiotemporal coverage of CCI SM, the data gap remains a major challenge that hampers its further application. Here, we select the daily combined microwave products version 4.5, with a spatial resolution of 0.25°. The inconsistent data in the CCI combined SM ~~is~~ are filtered using the quality flag variable.

A variety of Moderate Resolution Imaging Spectroradiometer (MODIS) products are collected, including the ~~0.05°~~ daily LST (MYD11C1), ~~the 0.05°~~ 16-day composite albedo (MCD43C3)~~, the 0.05° 16-day composite~~ and vegetation indices, i.e., normalized difference vegetation index (NDVI) and enhanced vegetation index (EVI), and the 8-day composite ~~500-meter~~ leaf area index (LAI) (MCD15A2H). All these datasets are collected at MODIS 6 collection. We calculate the diurnal temperature range (DTR) by subtracting the night LST from daytime LST. The NDVI and EVI are averagely obtained from the two products: MOD13C1 and MYD13C1. All the selected products are screened out using the quality variables to maintain only the available pixels with good quality. We also collect the 0.05° annual land cover product (MCD12Q1) for quality control of CCI SM.

~~Topography may be related to the spatial distribution of soil moisture.~~ We use the Digital Elevation Model (DEM) dataset provided by NASA's Shuttle Radar Topography Mission (SRTM) (Van Zyl, 2001) to retrieve several relevant topographic metrics~~. SRTM dataset has been extensively employed (Van Zyl, 2001), and it has an original spatial resolution of 90 m. The~~, including slope, aspect, and the topographic position index (TPI)~~ (Guisan et al., 1999) are used~~. The TPI is calculated by subtracting the focal grid elevation ~~with and~~ from the mean elevation of the eight surrounding grids. The TPI is potentially correlated better with surface variables such as snow depth and SM in comparison with the DEM (Cristea et al., 2017). Positive (negative) TPI values mean that the target grid is higher (lower) than the average of its surroundings.

Considering the low accuracy of satellite SM for snow-covered pixels, ~~the pixel~~ pixels that ~~has~~ have both daytime LST lower than 0 °C and ~~the~~ albedo higher than 0.3 are removed (Cui et al., 2020). We also remove ~~the pixel that occupies more than 20% of the~~ pixels for which a water body accounts of more than 20% of the total area. To overcome the spatial resolution differences among the diverse available~~ products~~, all datasets are resampled to 0.25° spatial resolution by averaging the pixel values.

~~2.1.2~~3.1.2 Reanalysis dataset and land surface model ~~dataset~~outputs

We collect the soil moisture data from ERA5, one global atmospheric reanalysis dataset released by the ECMWF (Balsamo et al., 2015). The data assimilation system used for ERA5 is the ECMWF Integrated Forecast System (IFS), and the meteorological forcing for retrieving soil moisture is from the ERA atmospheric reanalysis. ~~It can provide soils at four soil depths (0–7, 7–28, 28–100, and 100–289 cm).~~ Here we select the daily averaged SM from the first soil layer (0–7 cm) to match with satellite CCI SM.

Daily potential evapotranspiration (PET) and surface soil moisture (0–15 cm) is collected from the Global Land-surface Evaporation Amsterdam Methodology (GLEAM) dataset. GLEAM is based on a general land surface model that focuses on soil moisture and evapotranspiration (Miralles et al., 2011). PET in the GLEAM is calculated with the Priestley–Taylor formula based on multiple reanalysis datasets~~.~~, while the soil moisture is calculated with a soil-water module based on water cycle balance.

~~The soil moisture is calculated with a soil-water module based on water cycle balance.~~

Four meteorology variables, ~~are obtained from the Chinese regional ground meteorological dataset. They are~~ i.e., precipitation, air temperature, solar radiation~~,~~ and wind, are obtained from the China Meteorological Forcing Dataset. This dataset ~~has a temporal resolution of 3-hourly and a spatial resolution of 0.1°~~ is generated through fusion of in situ station data, remote sensing products, and reanalysis datasets (He et al., 2020). Considering the lag effect of precipitation on surface water dynamics, we use the five-day antecedent precipitation (AP) to replace the daily precipitation (Wei et al., 2020). ~~Specifically, AP is estimated as follows:~~

$$\cancel{AP = \frac{\sum_{i=1-m}^{i=0} P_i}{m},} \quad \cancel{(1)}$$

~~where~~ $\cancel{P_i}$ ~~is the precipitation at the day~~ $\cancel{i}$; $\cancel{m}$ ~~is the number of prior days for AP owing the highest absolute correlation between SM and precipitation. Here we set~~ $\cancel{m}$ ~~as five based on the regional hydrological and climatic variability.~~

Three surface temperature sources are additionally collected for uncertainty analysis. Two sources are collected from the ERA5 and GLDAS ensemble model. Considering the model uncertainties caused by regional surface characteristics and climatic conditions, we simulate surface temperature and surface ~~soil moisture~~SM (0–10 cm) by implementing ~~one~~ a Noah model that is forced with meteorology variables from the Chinese regional ground meteorological dataset and the surface condition parameters from MODSI. This dataset is previously used in our work (Liu et al., 2020a; Liu et al., 2021b). ~~All these surface temperatures and soil moisture are resampled to a 0.25° grid.~~

~~2.1.3~~3.1.3 In situ measurements

A variety of spatially sparse in situ soil moisture measurements is collected to evaluate the accuracy of gap-filled SM. We collect in situ soil moisture observations at 39 sites obtained from the China Watershed Allied Telemetry Experimental Research (WATER) project and the Chinese Ecosystem Research Network (CERN). These validation stations are set up in a relatively large homogeneous area dominated by vegetation covers (cropland, woodland and grassland) or desert lands. In addition, 657 in situ soil moisture measurements ~~are collected from the Chinese agro-meteorological and ecological observation network. These stations are~~ covered by cropland~~. All these selected in situ soil moisture measurements have been previously used for validating the satellite-derived soil moisture in China, and their locations and information are displayed in Fig.1 and Table 2~~ are collected from the Chinese agro-meteorological and ecological observation network.

We also collect the dense in situ measurements at the Maqu soil moisture monitoring network. The Maqu network (33°30′–34°15′N, 101°38′–102°45′E) is located on the north-eastern border of the Tibetan Plateau (Fig. 1) (Dente et al., 2012). In this network, 20 sites are distributed over a uniform grassland cover, located in the large valley of the Yellow River. Maqu network has demonstrated strong capability in monitoring the spatial and temporal SM variability with high accuracy (Su et al., 2013; Wei et al., 2019). The locations and detailed information of all available sites~~ is summarized in Table 2~~ are displayed in Fig.1 and Table S2.

~~**Table 2 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., 2021a) (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., 2018a) (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)~~

~~3. Methods~~

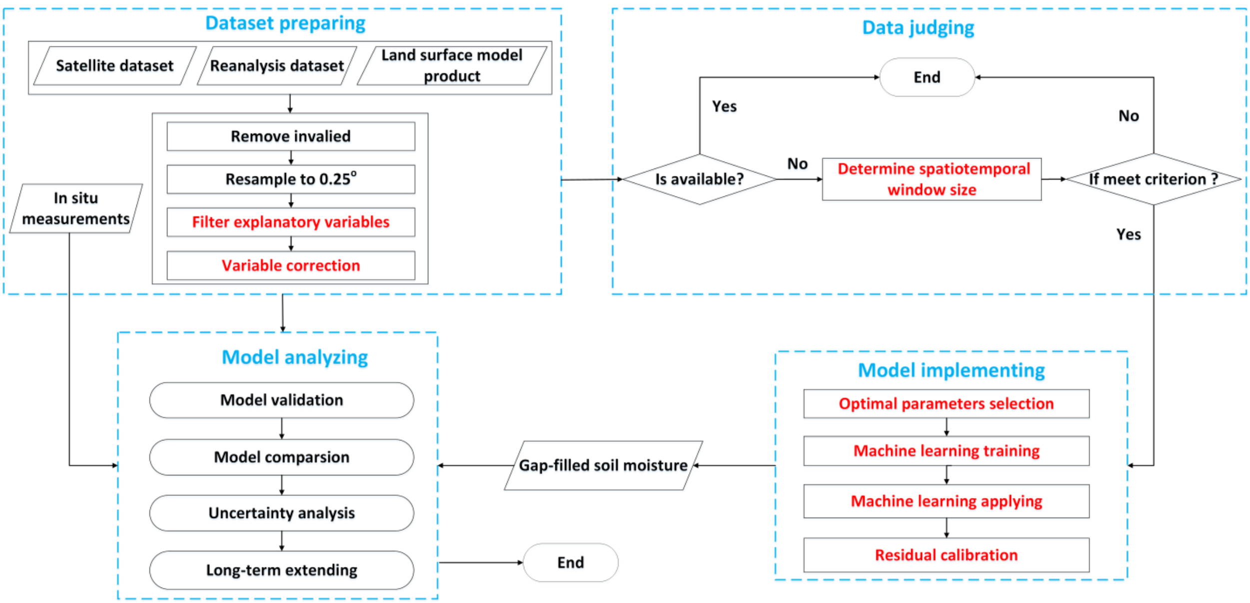
~~Our study aims to reconstruct the CCI SM data gaps for obtaining spatially continuous records. The basic idea beneath the proposed gap-filling approach is to efficiently depict the correlation between the SM records and the corresponding explanatory variables, which can be expressed as:~~

~~$SM = f(V_1, V_2, V_3 \cdots\cdots V_k) + \varepsilon,$~~ ~~(2)~~

~~$V_i \in R^{N,T},$~~ ~~(3)~~

~~where SM is the soil moisture, V_i is the corresponding explanatory vectors, and k is the number of the input variables. V_i can be a vector the sample number of which is decided by the spatial domain (N) and temporal domain (T). f is one function that can be either linear or nonlinear. ε represents the model residual. In machine learning ensemble, f represents a black box model that does not have one specific form.~~

~~Proposed methodology mainly involves the following steps: (i) using a regression subset selection model and a variable correction procedure to filter explanatory variables from the satellite observations and model-driven knowledge, and removing the systematic bias between them; (ii) applying a random forest algorithm to delineate the SM-explanatory variables correlation based on the available pixels identified with a spatiotemporal window search strategy, and then employing the established correlation to retrieve the unavailable SM pixels; and (iii) conducting a geographically weighted regression and Gaussian filtering to calibrate the model-derived residuals. Figure 2 shows the overall diagram of our work.~~



~~**Figure 2: The overall diagram of the proposed methodology. The main work carried out includes dataset preparing, data judging, model implementing, and model analyzing.**~~

~~3.1 Explanatory variables filtering~~

~~3.1.1 Explanatory variables~~3.1.4 Filter explanatory variables

Explanatory variables related to atmospheric, geophysical, ecological, and hydrological variables are conductive in capturing SM variability. ~~The main assumption beneath this 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.~~ The significance percentage produced ~~in~~ by the regression subset selection model (Fu et al., 2019; Liu et al., 2021a) is employed to measure the impacting probability of explanatory variables, ~~and the~~ where a high significance percentage indicates ~~the~~ strong ~~ability to depict SM~~ capability in depicting SM (details in Text S1). ~~To be specific, we successively (1) use a least-squares linear regression to check the potential relationships between SM and explanatory variables; (2) apply a stepwise regression to explore plenty of potential explanatory variables based on the Akaike Information Criterion (AIC); (3) exploit the best models from all variable combination to identify the important variables impacting SM; and (4) quantify the relative contributions of each explanatory variable to SM based on importance criterion.~~

We conduct the subset selection model analysis based on ~~the~~ a dataset ~~during~~ from 2005~~—~~ to 2015~~.~~, and 15 variables ~~are~~ were selected as input parameters, including seven surface environmental variables, i.e., Albedo, NDVI, EVI, LAI, DTR, PET and ERA SM, three elevation variables, i.e., TPI, aspect and slope, three climatic variables, i.e., AP, air temperature, wind, and two geographical factors, i.e., latitude and longitude. ~~Notice that all these~~ All the variables ~~can be~~are available from ~~a~~ reliable dataset at the continental scale. ~~Moreover, these variables have been reported previously (Cui et al., 2016; Liu et al., 2020b; Almendra-Martín et al., 2021) in robustly describing soil moisture.~~ Gaps presented in these variables were not considered further to avoid introducing additional errors.

As illustrated in Fig. 3(a), ~~Albedo~~albedo, NDVI, EVI, LAI, DTR, AP, PET, ERA SM, TPI and air temperature have the highest significant percentage in terms of correlating ~~to~~ with CCI SM. ~~Considering EVI and LAI have closely correlated to NDVI, and air temperature has closely correlated to DTR. EVI, LAI and air temperature are excluded for model application.~~ We excluded aspect, slope, wind, latitude, and longitude owing to their low correlations with SM. The EVI, NDVI, and air temperature were also not considered in further application because the EVI and LAI are closely correlated with NDVI, and air temperature is strongly correlated with DTR. All ~~these~~ the selected covariates are physically meaningful in depicting SM. Specifically, ~~Atmospheric~~ the atmospheric variables (i.e., precipitation, DTR and PET) are suitable to capture the temporal dynamics of SM~~.~~, ~~Topographic~~ and the topographic variables are included ~~to~~ both to depict the orographic effects and to recapture the spatial pattern of SM. ~~NDVI and Albedo (and DTR) are included owing to their impact on the formation of SM. Specifically,~~ DTR exhibits ~~a substantial~~ correlation with SM owing to its capacity in taking account ~~for~~ of land-atmosphere coupling. ERA surface moisture ~~is~~ was also included to reproduce satellite SM. ~~Despite uncertainties in the numerical model SM simulations, ERA SM provides the fewest dataset gaps. We further investigate the contribution of these variables to SM. As shown in Fig. 3(b) (and Fig. S1), the correlations between the CCI SM and the selected variables are relatively high across the whole of China. This indicates the feasibility of the selected variables in modeling SM. In addition, since these variables are derived from optical remote sensing, reanalysis dataset and land surface model products, they have the potential to extend to large regions due to high availability (Fig. 3(c)).~~

To verify the results based on the regression subset selection model, we employed the permutation feature importance to measure the relative importance of each predictor variable. Consistent patterns between the significance percentage and permutation importance further indicate the feasibility of the selected variables in modelling SM. Additionally, because these variables are derived from optical remote sensing, reanalysis datasets, and land surface model products, they have potential for extension to large regions owing to their high availability (Fig. 3(b)).

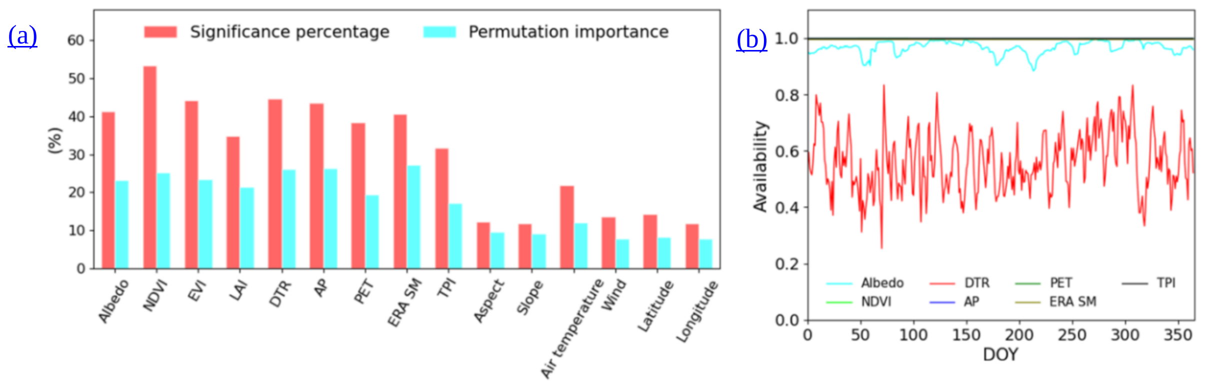


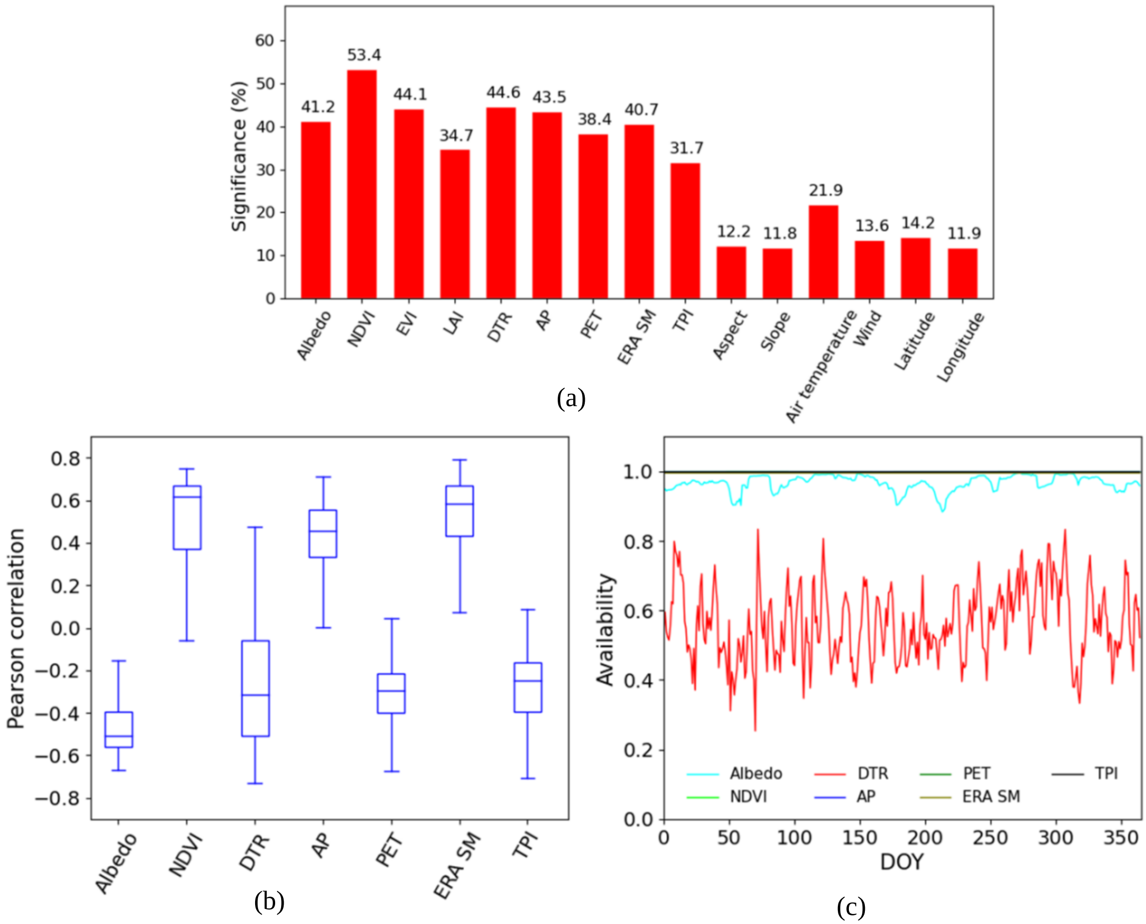
Figure 3: The correlation and availability of dataset used. (a) The significance percentage and permutation importance of the selected variables in correlation to CCI SM. (b) The availability of the selected variables.

~~3.1.2~~3.1.5 Variable correction

Systematic biases are unavoidable in reanalysis datasets and land surface model outputs, and these biases can be propagated in dynamic modeling. Accordingly, bias correction is required prior to the gap-filling procedure to ensure a consistent simulated output. Specifically, ~~To~~ to make the modeled values (i.e., ERA SM) comparable ~~to~~ with the satellite observations (i.e., ESA CCI SM), ~~it is necessary to remove the systematic bias between them. Here~~ we ~~use one~~ used a correction procedure ~~(Long et al., 2020; Zhang et al., 2021d), which~~ that primarily combines ~~the~~ a variance scaling algorithm and ~~the~~ a linear scaling algorithm (Long et al., 2020; Zhang et al., 2021c). The used procedure can be illustrated with the following equations:

$$\begin{cases} SM_{c1} = SM_{ERA}(t_{av}) - \mu\big(SM_{ERA}(t_{av})\big) + \mu\big(SM_{ESA}(t_{av})\big) \\ SM_c = \mu(SM_{c1}) + (SM_{c1} - \mu(SM_{c1})).\sigma\big(SM_{ESA}(t_{av})\big)/\sigma\Big(SM_{c1} - \mu\big(SM_{p1}\big)\Big) \end{cases}, \qquad (~~4~~3)$$

where SM_{ERA} is the raw ERA SM time series of the target grid pixel; t_{av} is time series in which pixels in the object grid are available; SM_{ESA} is the ESA SM of the grid; μ and σ are the mean value and the standard deviation, respectively. SM_c is the corrected ERA SM that is assumed to have a spatial pattern (i.e., consistent means and standard deviations) with the CCI SM. In our study, a dataset ~~regarding the~~ comprising time series ~~of~~ from 2005~~—~~ to 2015 ~~is~~ was used to conduct the correction procedure ~~for guaranteeing enough~~ to guarantee sufficient samples. ~~The examples~~ Examples illustrating the performance of the ERA SM correction can be found in Fig. ~~S2~~S1. Despite being conducted on SM~~ parameters here~~, this calibration ~~method~~ ~~can~~ procedure could be ~~implemented~~ applied to other parameters (e.g., DTR) when replaced with numerical model outputs.



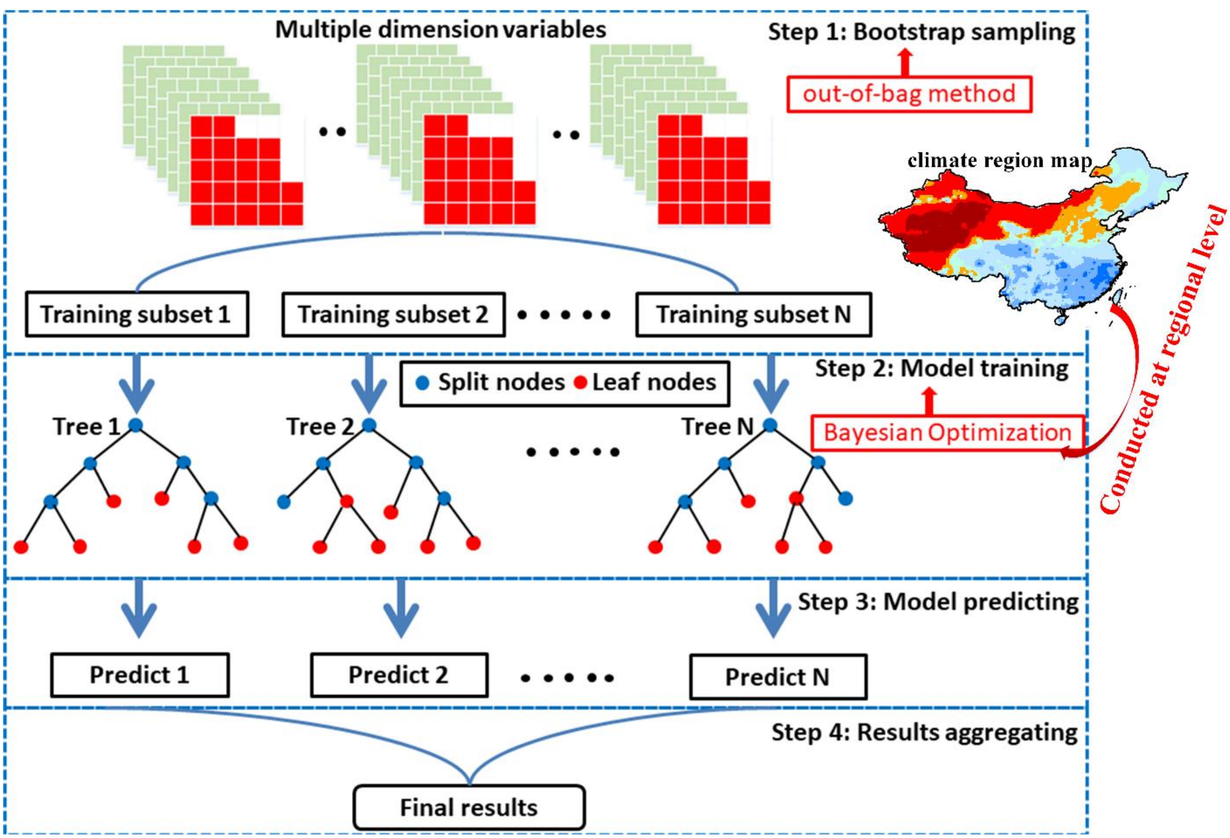
~~**Figure 3: The correlation and availability of dataset used. (a) The significance percentage of the selected variables in correlation to CCI SM. (b) The Pearson correlation between the selected variables and CCI SM. (c) The availability of the selected variables.**~~

3.2 ~~Random Forest regression~~ Model implementation

3.2.1 Machine learning regression

Despite being easy to implement and requiring less computational resources, ~~the~~ traditional regression-based methods ~~cannot provide full probability density functions for evaluating system performance. Ensemble learning approaches may overcome this issue by aggregating results from multiple models~~ such as generalized linear models and multivariate regression splines generally insufficiently consider the probability density functions in assessing model performance. Machine learning approaches could ~~achieve better performance owing to their capacity in reducing overfitting chances and quantifying the accompanying uncertainty~~be much more flexible than conventional parametric models owing to their ability to handle nonlinear relationships and complex interactions. Among the machine various learning models, the random forest (RF) algorithm, acting as an enhanced decision tree model, is ~~one~~ an effective and powerful tool in interpreting earth variables~~, acting as an enhanced decision tree model~~ (Belgiu and Drăguţ, 2016). As illustrated in Fig. 4(a), RF is a hierarchical tree diagram~~, which~~ that is based on a nonparametric strategy and ~~therefore is feasible to add layer categories~~ has the capacity to add a variety of parameter layers into the model (Breiman, 2001). This decision tree model is composed of many nodes and edges within each tree structure, mainly including two types of nodes: split nodes and leaf nodes. The split node is related to a test function that is employed to split the input data, whereas the leaf node is associated with the final decision. Unlike the standard decision tree model that relied on the whole data set, RF trains each tree on bootstrap resamples. This model only considers the randomly selected variables rather than the total variables. By this means, the outcome is decided by ~~one~~ a majority voting or averaging strategy.

In this study, the RF model is implemented using the function 'RF Regressor' from the Python Library (Shahriari et al., 2016). Specifically, the built-in functions are used to assess the importance of each covariate by using the out-of-bag samples. We use the 'Bayesian Optimization' module (http://rmcantin.github.io/bayesopt/html/bopttheory.html) to select the best hyperparameters in driving RF algorithm. Four critical parameters deciding the RF algorithm include the number of trees (n_estimators), the maximum tree depth (max_depth), the minimum number of samples for splitting an internal node (min_samples_split), and the number of features (max_features). For each specific climate region, the Bayesian optimization process is carried out within 20 iterations to optimal parameters. The training procedure is mainly based on the dataset covering 2003—2008. Optimal parameters in the seven climate regions are listed in Table ~~3~~S3.



~~**Figure 4: The diagram of the random forest model implemented for a multidimensional dataset. Stage 1: The out-of-bag sampling method is used to sample time series dataset. Stage 2: The independent decision trees for model training are built based on the optimal parameters from the Bayesian Optimization module. This process is conducted at the climate regional level. Stage 3: The predicted values are produced from each bootstrap tree. Stage 4: The outcome is obtained by averaging earlier results from multiple trees.**~~

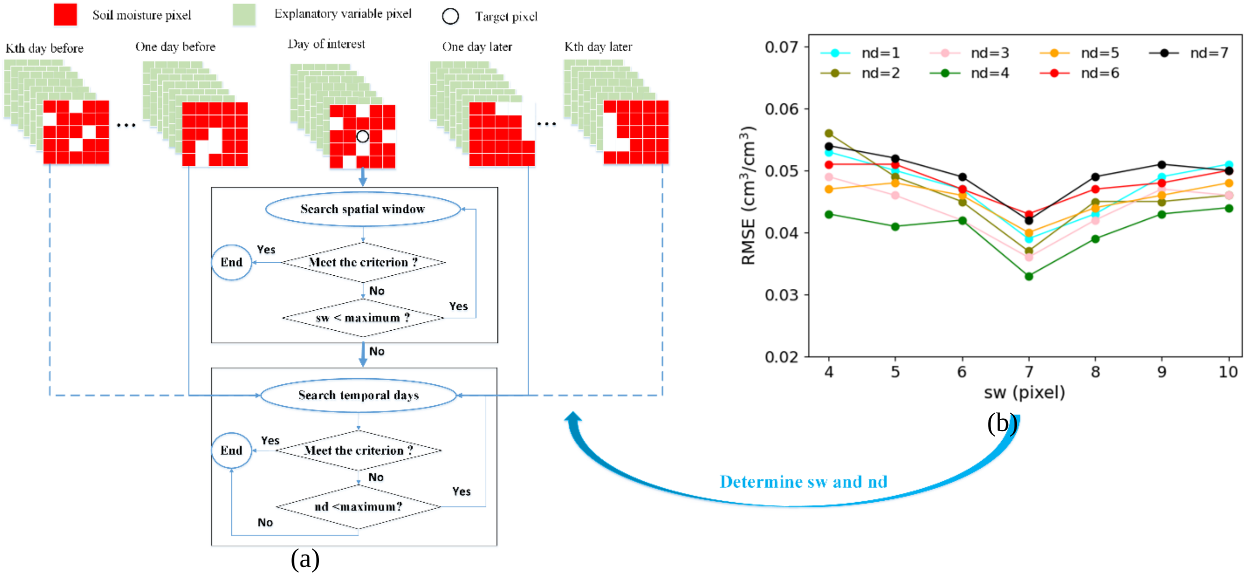
~~**Table 3 Optimal parameters regarding seven climate regions**~~

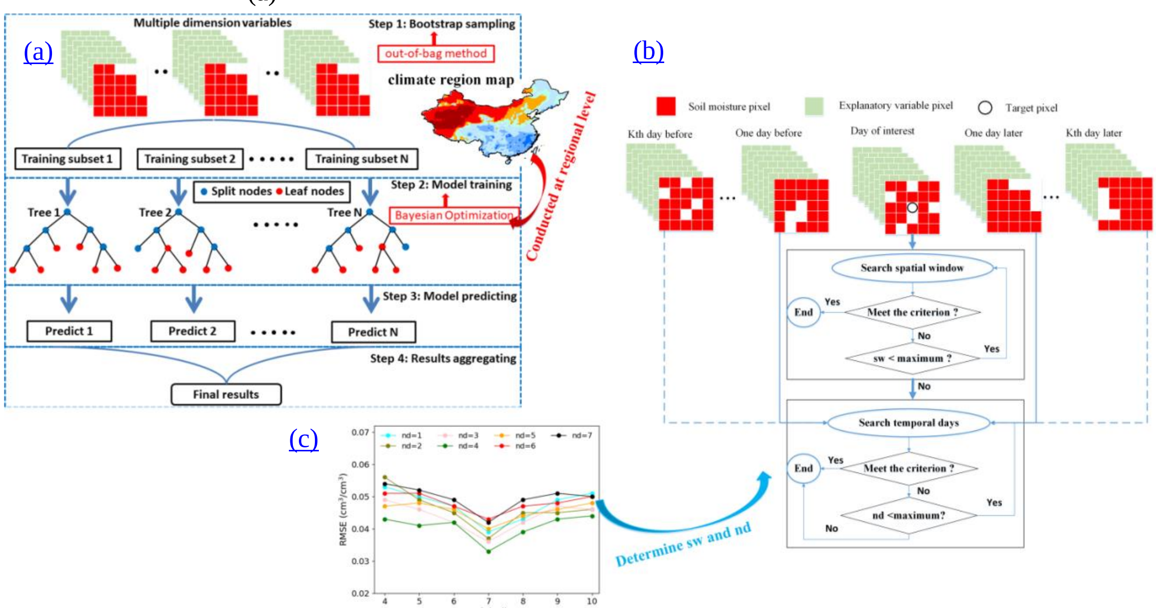
~~Climate region~~	~~n_estimators~~	~~max_depth~~	~~min_samples_split~~	~~max_features~~
~~Arid~~	~~69~~	~~11~~	~~8~~	~~0.12~~
~~Semi-arid~~	~~80~~	~~18~~	~~9~~	~~0.16~~
~~Aird/semi-wet~~	~~47~~	~~9~~	~~5~~	~~0.31~~
~~Wet/semi-arid~~	~~36~~	~~10~~	~~3~~	~~0.25~~
~~Wet~~	~~52~~	~~15~~	~~11~~	~~0.16~~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

~~3.3 Spatiotemporal strategy~~ 3.2.2 Identify spatiotemporal window

One critical issue ~~of~~ relate to the machine learning model is how to efficiently explore the informative covariates. Here, we use ~~one~~ a spatiotemporal strategy to ~~most~~ capture the spatial and temporal SM and the related covariate dynamics. Our strategy primarily relies on the available pixels within ~~one~~ a regional subset, ~~thus including more interest pixels to participating regression~~ thereby allowing more pixels of interest to participate in the regression. Figure ~~5(a)~~4(b) provides the diagram of the spatiotemporal window search strategy.

~~One~~ An adaptive strategy is employed to determine the optimal spatiotemporal window size. Two critical variables are adopted to identify the window size, i.e., the size of the spatial window (sw) and the number of temporal days (nd). To find the optimal sw and nd, we continually increase the value of sw and nd from the initial values until the samples participating for regression meet the criterion, i.e., the number of available pixels within the searched window should be no less than eight times of the participating explanatory variables (i.e., seven) (Svetnik et al., 2003; Liu et al., 2020a). Here an initial sw is set to 5 and an initial nd is set to 1. Considering that a fraction of gaps occur in the satellite dataset (e.g., LST and albedo) and the optimal window may not exist, the maximum values of sw and nd are introduced to terminate this process. ~~One~~ A sensitivity analysis is conducted with the independent dataset to select the two maximum values. Specifically, we conduct ~~one~~ a cross validation during 2003—2008 to evaluate the accuracy of the gap-filling model. The increasing maximum nd from 1 to 7 with intervals of length 1 is tested, and the maximum sw is tested from 4 to 10 with intervals of length 1. The values that yield the lowest RMSE (Fig. ~~5(b)~~4(c)) are selected, and finally, we set maximum sw to 7 and the maximum nd to 4. Note that we also conduct sensitivity analysis for each climate region and find no substantial differences in the resulting optimal values of two parameters among seven climate regions. This is probably because this sensitivity analysis is more reliant on model structure rather than sample characteristics.





~~**Figure 5: (a) The diagram of spatiotemporal window determination strategy for random forest regression. (b) The results of the sensitive analysis regarding two maximum values for terminating the searching process.**~~

Figure 4: (a) The diagram of the random forest model implemented for a multidimensional dataset. (b) The diagram of spatiotemporal window determination strategy for random forest regression. (c) The results of the sensitive analysis regarding two maximum values, i.e., the size of the spatial window (sw) and the number of temporal days (nd), for terminating the searching process.

~~3.4~~3.2.3 Residuals calibration

Considering that the machine learning model ~~may~~ might not fully account for the variability in ~~soil moisture~~SM, the original reconstruction needs to be calibrated~~ with the corresponding model residuals. The calibration procedure can correct the bias resulting from neglectful variables~~, which can potentially remove the bias resulting from neglected variables such as those are excluded for model establishment (Zhu et al., 2012; Liu et al., 2020a). ~~This is very critical considering the relatively less spatial representative of soil moisture within the 25 km scale.~~ In practice, we add the interpolated residuals to the original reconstructions.

~~A GWR model~~ The geographically weighted regression (GWR) model, which is an extension of the traditional linear regression model (Li et al., 2017), is applied to interpolate the RF-derived residuals. This procedure is based on the samples within the searched window for each target pixel. The model residual (ε_j) derived from Eq. (2) can be described using the explanatory variables as follows:

$$\varepsilon_j = \beta_0(u_j, v_j) + \sum_{i=0}^{k} \beta_i(u_j, v_j) X_{ij}, \qquad (\cancel{5}4)$$

where $\beta_0(u_j, v_j)$ and $\beta_i(u_j, v_j)$ are the regression coefficients estimated at the jth pixel, and (u_j, v_j) are the coordinates. The regression coefficients can be estimated using the observations within the self-adaptive searched window as follows:

$$\begin{cases} \hat{\beta}(u_j, v_j) = (X^T(W(u_j, v_j))X)^{-1} X^T W(u_j, v_j) Y \\ \qquad\qquad w_{ij} = [1 - (d_{ij}/b)^2]^2 \end{cases}, \qquad (\cancel{6}5)$$

where $\hat{\beta}(u_j, v_j)$ is the coefficient matrix composed of coefficients from each explanatory variable; X and Y are the explanatory variable matrix and the dependent variable (i.e., SM) vector, respectively. Her latitude, longitude and seven explanatory variables selected ~~in section 3.1.1~~ are used to implement the GWR model. $W(u_j, v_j)$ is the weight matrix composed of w_{ij}, d_{ij} is the Euclidean distance between the observation ith and the jth point, a and b is the window radius. Before adding to the original reconstruction, the GWR interpolated residual is further smoothed with a normalized k × k Gaussian filter with a standard deviation of σ. This procedure can remove the grid-like artifacts that extensively exist in statistical model outcomes. Base on the optimization procedure (Sismanidis et al., 2021; Liu et al., 2019), we set k = 5 and σ =1.5.

~~3.5 Accuracy evaluation~~3.3 Model analysis

3.3.1 Model validation

Model validation was conducted using data from 2009 when sufficient number of ground measurements were collected. The top layer ~~(Table 2)~~ SM measurements from in situ stations ~~are~~were first used to evaluate the accuracy of the reconstructed results. Considering the scale mismatch between the sparse distribution of in situ ~~station~~ stations and CCI SM product (~25 km), we use the Disaggregation based on Physical And Theoretical scale Change (DISPATCH) model (Merlin et al., 2012) to disaggregate the 0.25° reconstructions to 1 km resolution. ~~As one typical SM disaggregation model, DISPATCH has been extensively applied in current studies (Molero et al., 2016; Song et al., 2021). In DISPATCH, SM can be depicted at different spatial scales by linking to LST and NDVI through soil evaporative efficiency.~~ Detailed descriptions regarding this disaggregation method can be found in Supplementary Text ~~S1~~S2.

Evaluating the gap-filled SM with in situ measurements is supposed to produce biases that can be caused by scale mismatching and disaggregation model performance. To account for this, holdout cross validation with 10 replicates was performed in 2009 to evaluate the model accuracy. ~~In addition to field measurements, one holdout cross-validation with ten replicates is conducted to comprehensively evaluate the model performance.~~ For each replicate,~~ we separate the dataset (CCI SM and explanatory variables) into the training (90%) and the test subsets (10%)~~ we randomly held out 10% of the pixels, that is manually introducing gaps for these pixels, and trained the model with the remaining 90% of the dataset. Specifically, the pixels during all periods were first rearranged into a time series and then 10% of them were dropped in each replicate. After the gap-filled SM series ~~are produced with~~ of hold-out pixels were reconstructed from the training set, they ~~will be~~were validated~~ with the test set~~ against the original SM. ~~The above cross-validation is also applied to conduct model comparison and uncertainty analysis.~~

To reveal the physical plausibility of gap-filled SM, we paid particular attention to the evaluation of gap-filling SM under extremely dry conditions. Extreme drought is defined based on meteorological condition, that is, the Palmer Drought Severity Index (PDSI) of less than −2 over 8 consecutive months or longer (Fig. S2).

The statistics used for the model accuracy assessment include the coefficient of determination (R^2), the root mean square error (RMSE), the mean absolute error (MAE), the average error bias (BIAS), and the unbiased RMSE (ubRMSE). In addition, Nash-Sutcliffe Efficiency (NSE) is used to measure the overall performance of the proposed model. All these metrics have been extensively used for evaluating satellite SM.~~, and they can be described as follows~~:

$$\cancel{R^2 = 1 - \frac{\sum_i^k (SM_i - \widehat{SM}_i)}{\sum_i^k (SM_i - \overline{SM})},} \tag{7}$$

$$\cancel{RMSE = \sqrt{\frac{\sum_i^k (SM_i - \widehat{SM}_i)^2}{k}},} \tag{8}$$

$$\cancel{MAE = \frac{\sum_i^k |SM_i - \widehat{SM}_i|}{k},} \tag{9}$$

$$\cancel{BIAS = \frac{\sum_i^k (\widehat{SM}_i - SM_i)}{k},} \tag{10}$$

$$\cancel{ubRMSE = \sqrt{RMSE^2 - BIAS^2},} \tag{11}$$

~~where SM_i is the reconstructed soil moisture of the ith sample, and $\widehat{SM}_i$ is the corresponding reference (or field) value, k is the number of samples.~~

3.3.2 Model comparison

The proposed method was compared against four extensively used models that adopt the same explanatory variables and spatiotemporal window search strategy. The first one is the conventional multiple linear regression (MLR) approach. Three typical machine learning approaches, i.e., Extreme gradient boost (XGB), Support vector machine (SVM) and Artificial Neural Network (ANN), are also used for comparison. Detailed descriptions of four available models can be found in supplementary Text S3.

3.3.3 Uncertainty analysis

Considering the criticality of explanatory variables in simulating SM, uncertainty analyses regarding these selected variables were conducted. We first investigated the accuracy of the reconstruction model that excludes one participating variable. Given the critical importance of satellite-derived DTR and the severe issues of missing data in satellite-observed LST products, we further investigated the substitution performance of other surface temperature sources in reconstructing SM, i.e., i.e., Noah, ERA and GLDAS. This analysis was conducted by focusing on two regions (in Fig. 1) that have sufficient data sources to support our experiments (Liu et al., 2020a; Liu et al., 2021b): one region is in northern China covering mostly arid and semi-arid areas, while the other region is in southern China covering mostly wet areas.
Since the reanalysis SM is a vital input in our approach, we also compare it with the other two products to evaluate the feasibility of ERA data in reconstructing CCI SM. GLEAM and Noah surface SM are respectively employed to replace the ERA SM while other explanatory variables keep the rest the same.

3.3.4 Long-term extension

The available dataset forcing for our model has a long record, indicating potential for modelling long-term SM products. To verify this, the proposed gap-filling method was further extended to the long-term ECA CCI SM databases of 2005—2015. We also investigated the trend of the SM series during this period, which was obtained via Sen's slope and M-K significance analysis (Li et al., 2021c; Liu et al., 2021a). The trends from the reconstructed SM series were also compared with those from the original CCI SM, which were evaluated against in situ measurements.

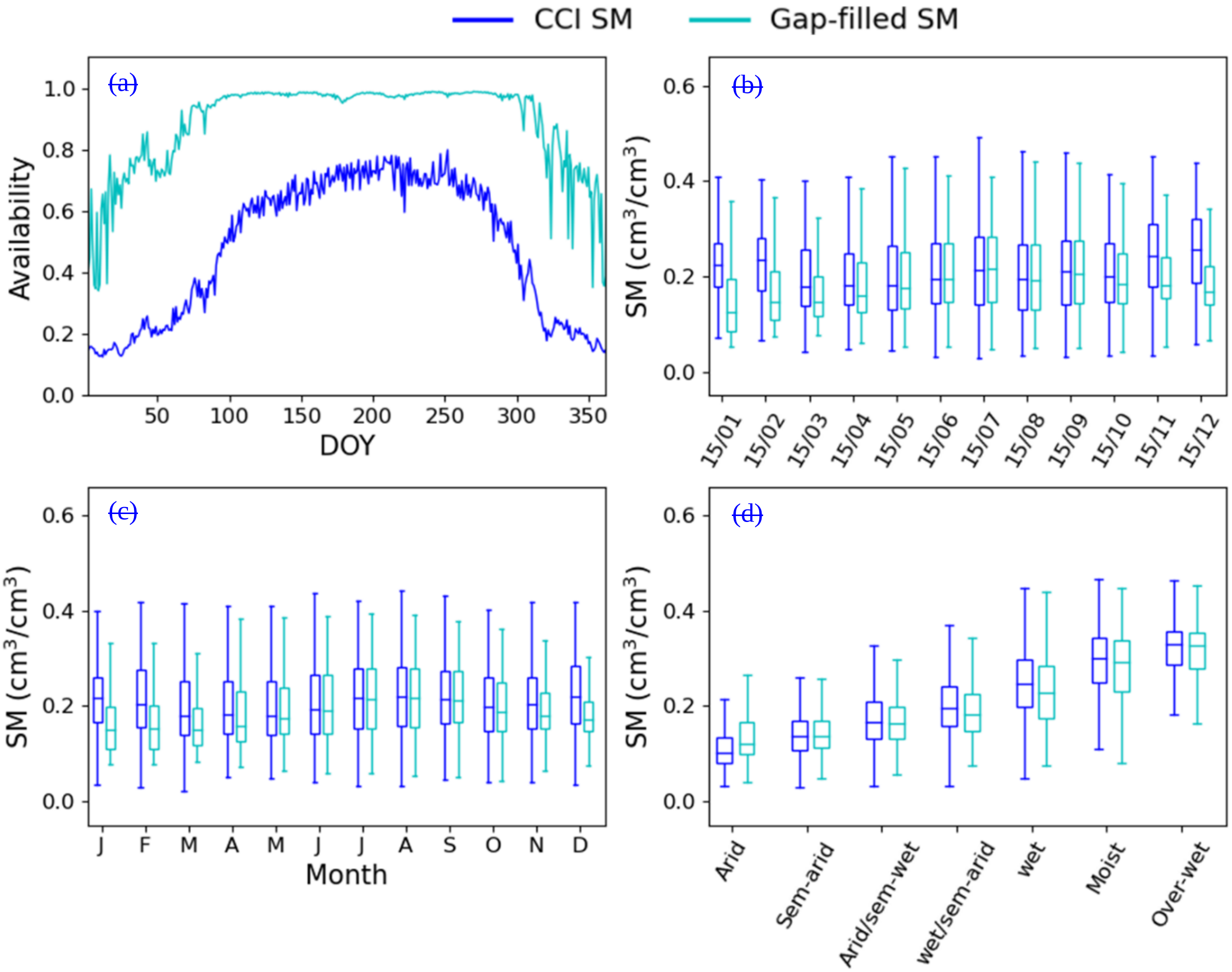
4. Results and discussions

4.1 Spatiotemporal patterns

The spatiotemporal pattern of the original daily CCI SM and the corresponding gap-filled dataset in 2009 is first checked. As shown in Fig. ~~6~~5(a) (and (Fig. S3)), a considerably large gap occurs in the original CCI SM, and this gap ~~issue~~ problem is ~~heavier~~ greater in ~~the~~ winter~~ season~~. We reconstruct the contaminated SM pixels using the spatiotemporal ~~random forest~~RF model. ~~It's observed that most~~ Most of the contaminated pixels (more than 85%) are reconstructed. Relatively fewer missing pixels are gap-filled in ~~the~~ winter ~~season~~ in comparison ~~to~~ with other seasons, primarily ~~relating to the heavy missing issues during this time~~ because of the heavy contamination of clear pixels caused by frequent occurrence of cloud during this period. It means that the learning capacity of the spatiotemporal machine learning method is constrained when encountering limited satellite observations.

Figure ~~6~~5(b) shows the boxplot of original versus gap-filled SM on ~~the~~ selected days in 2009. ~~Relative minor conformity~~ Conformity exists between the original and reconstructed SM for most days. A ~~consistent~~similar pattern ~~between them is also observed on the monthly-average SM~~ in variance and magnitude is also observed for the SM of the monthly average and the selected days, as illustrated in Fig. ~~6(c).~~ 5(c)~~.~~; that is, large ~~Large~~ differences ~~occur~~ occurs in ~~the~~ winter and spring ~~seasons~~. This can be attributed to the fact that the original CCI SM provides ~~less~~ fewer training data from October to May of the following year. ~~In addition~~ Additionally, the distribution of CCI SM is more uneven in this period, which ~~may~~ might reduce ~~the~~ model performance ~~due~~ owing to the limited representation of training samples (Stroud et al., 2001).

~~We further check the model performance regarding different climate regions. Figure 6(d) demonstrates one~~ In terms of different climate regions, minor discrepancy is evident between the original and the reconstructed SM (Figure 5(d)), with ~~the~~ bias in the median SM values of less than 8%. ~~This~~ It means that the reconstructed SM ~~owns a~~ has strong variation ~~delineating~~ depicting capacity. ~~A small~~Small overestimation occurs ~~for the~~in arid regions, which originally have less soil water storage.



CCI SM
Gap-filled SM
~~(a)~~
Availability
DOY
~~(b)~~
SM (cm³/cm³)
15/01 15/02 15/03 15/04 15/05 15/06 15/07 15/08 15/09 15/10 15/11 15/12
~~(c)~~
SM (cm³/cm³)
J F M A M J J A S O N D
Month
~~(d)~~
SM (cm³/cm³)
Arid
Sem-arid
Arid/sem-wet
wet/sem-arid
wet
Moist
Over-wet

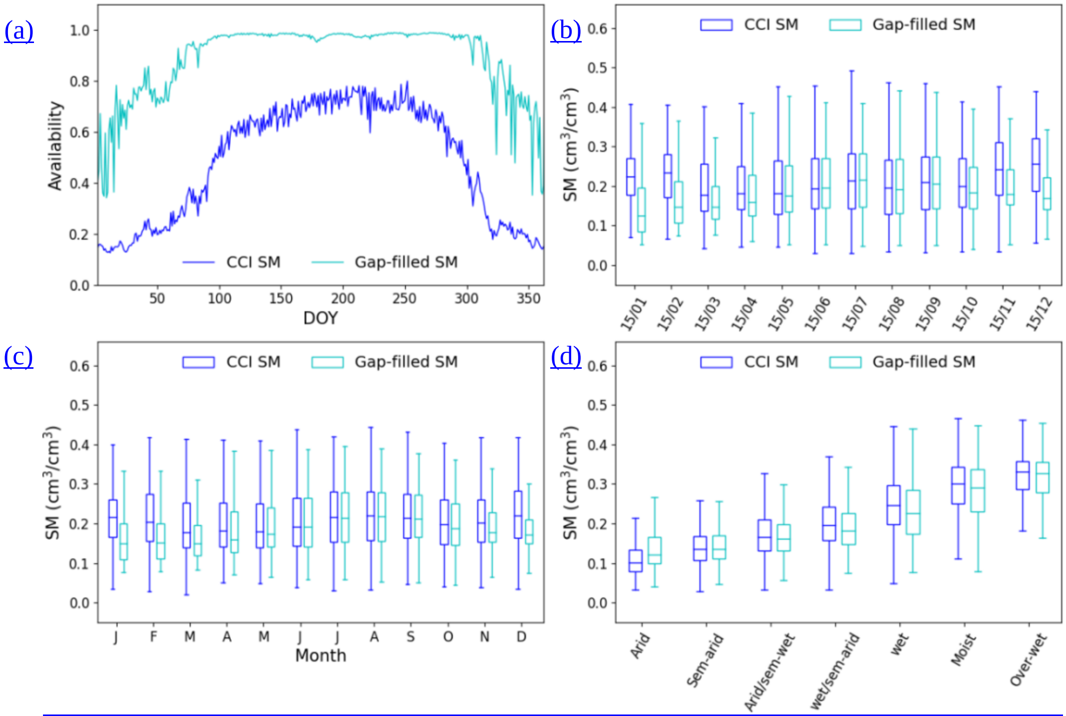
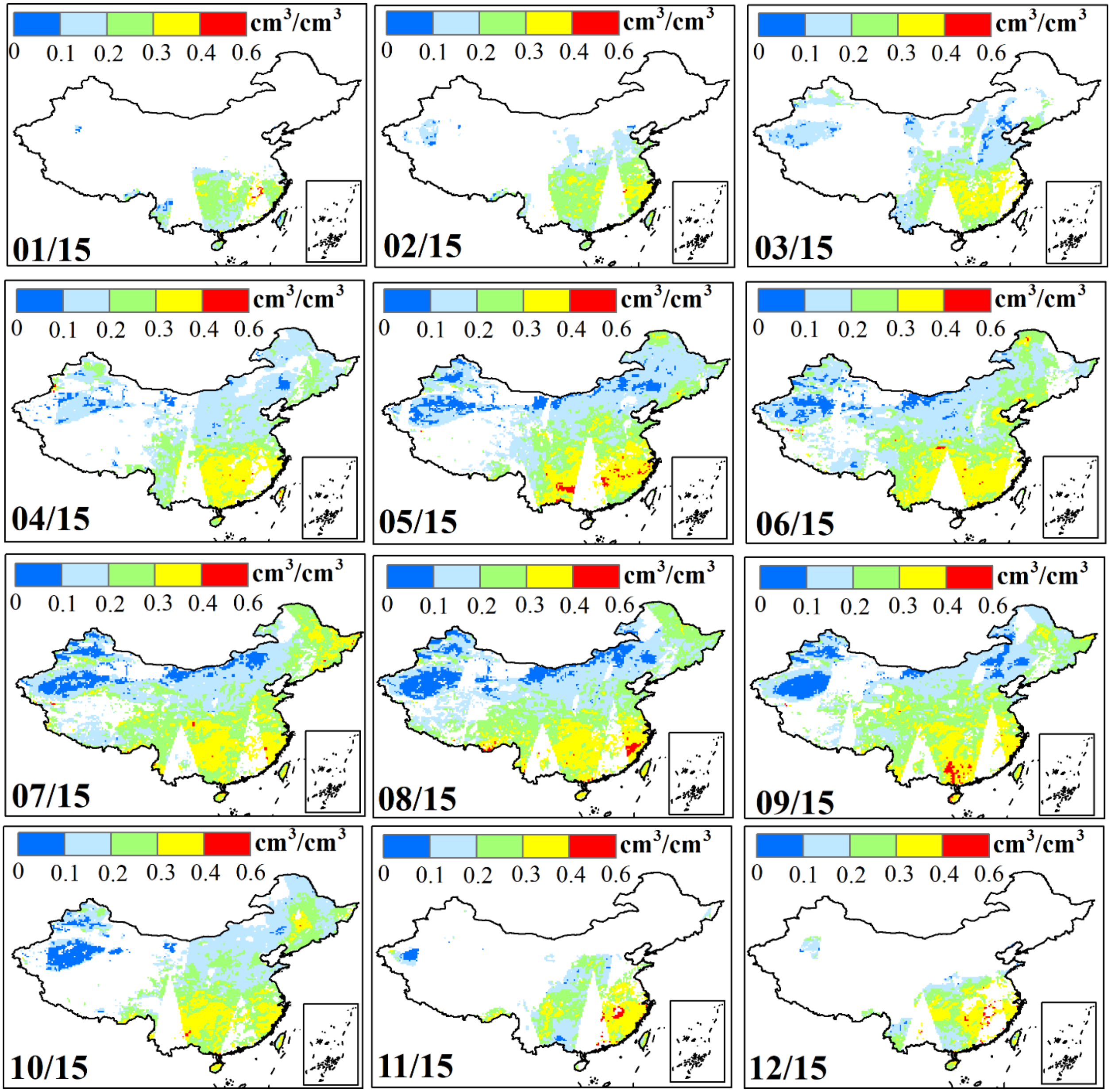
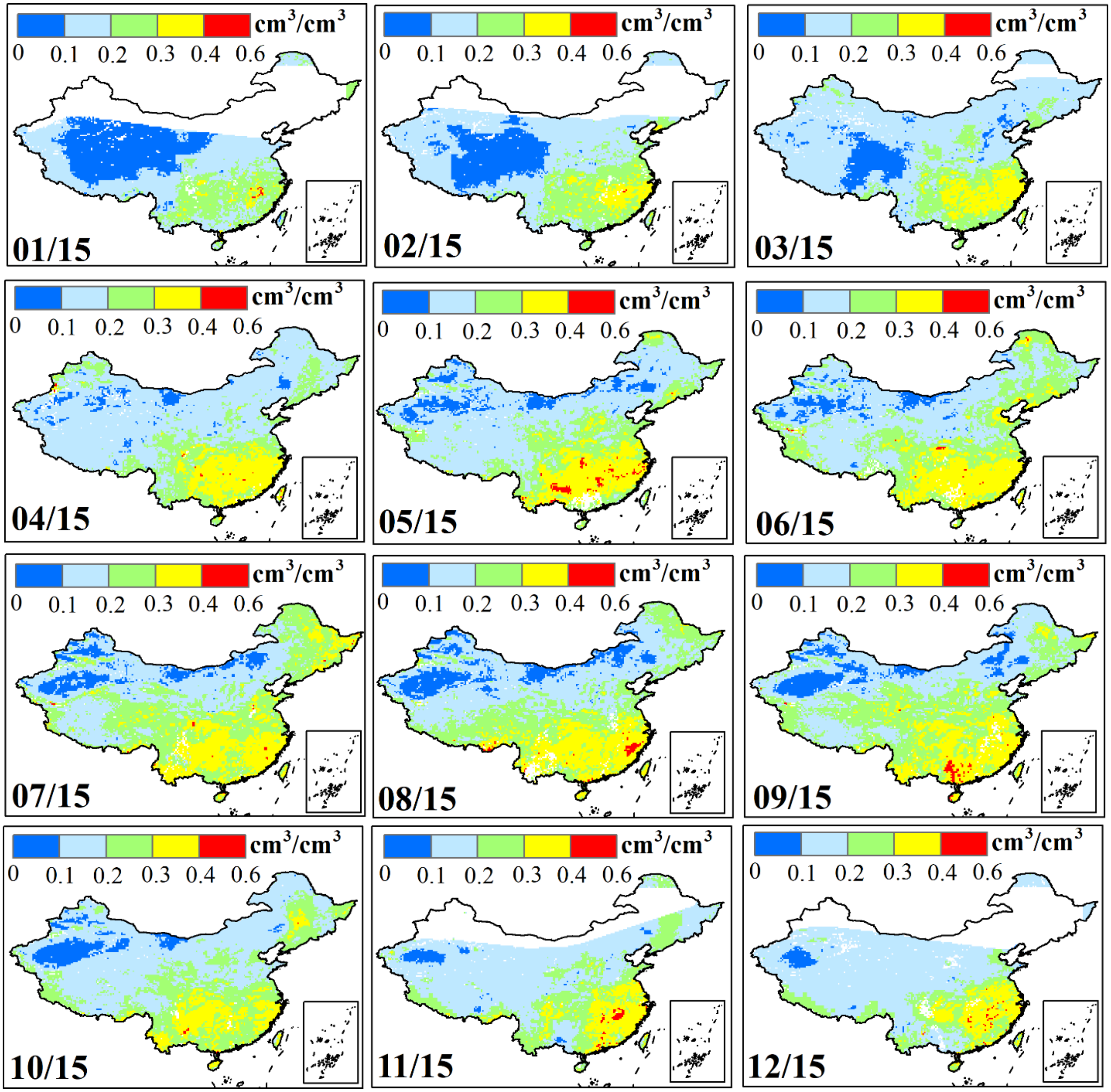


Figure ~~6~~5: The comparison between CCI dataset and gap-filled SM in 2009. (a) The plots of the availability of CCI dataset and gap-filled SM. (b) The boxplot of the CCI dataset and gap-filled SM on the selected days. (c) The boxplot of month-average CCI and gap-filled SM. (d) The boxplot of raw and gap-filled SM regarding seven climate regions.

Figure ~~7~~ 6 exhibits the spatial distributions of the original CCI SM and the reconstructed SM on ~~the~~ selected days in 2009. ~~It's observed that the~~The humid regions are mostly concentrated in southern China ~~that is~~ adjacent to the west coast of the western Pacific, whereas the dry regions are mainly distributed in ~~the~~ northern and western parts of China. A considerable fraction of contaminated pixels is observed on the selected days, and this contamination ~~issue~~ is severe in ~~the~~ winter season and in mountainous areas (e.g., Tibet Plateau and Mongolian Plateau). ~~The spatial distributions of the reconstructed SM corresponding to the selected days are illustrated in Fig. 8.~~ Almost all the contaminated pixels from March to October are reconstructed; meanwhile, the proposed model reconstructs the most contaminated pixels for the remaining months. Owing to the additional valid values provided by gap-filled pixels, more spatial ~~variations are delineated~~ variation is depicted in the reconstructed SM images. ~~The missing~~Missing pixels still occur in the reconstructed SM images especially in the cold seasons. This can be related to the fact that the surface temperature, ET, and precipitation are more connected in the warm season through energy balance considerations and atmospheric circulation. Some of these ~~invalidate~~ invalid pixels correspond to the ~~snow~~ snow- and water-covered regions that have been ~~beforehand~~ removed beforehand. Because missing earth data are to a large extent not at random, statistical measures of comparative analysis among them tends to produce bias (Bessenbacher et al., 2022b). To account for this, paired histograms of two datasets are compared to explore the value distribution properties. The histograms show the gap-filled dataset does not impact the SM distribution in warm seasons, that is, in agreement with the CCI dataset. There is also noticeable bias in cold seasons, especially in the very low range of SM.



~~Figure 7: The spatial distributions of raw CCI SM on the 15th of each month in 2009.~~



~~**Figure 8: The spatial distributions of gap-filled CCI SM on the 15th of each month in 2009.**~~

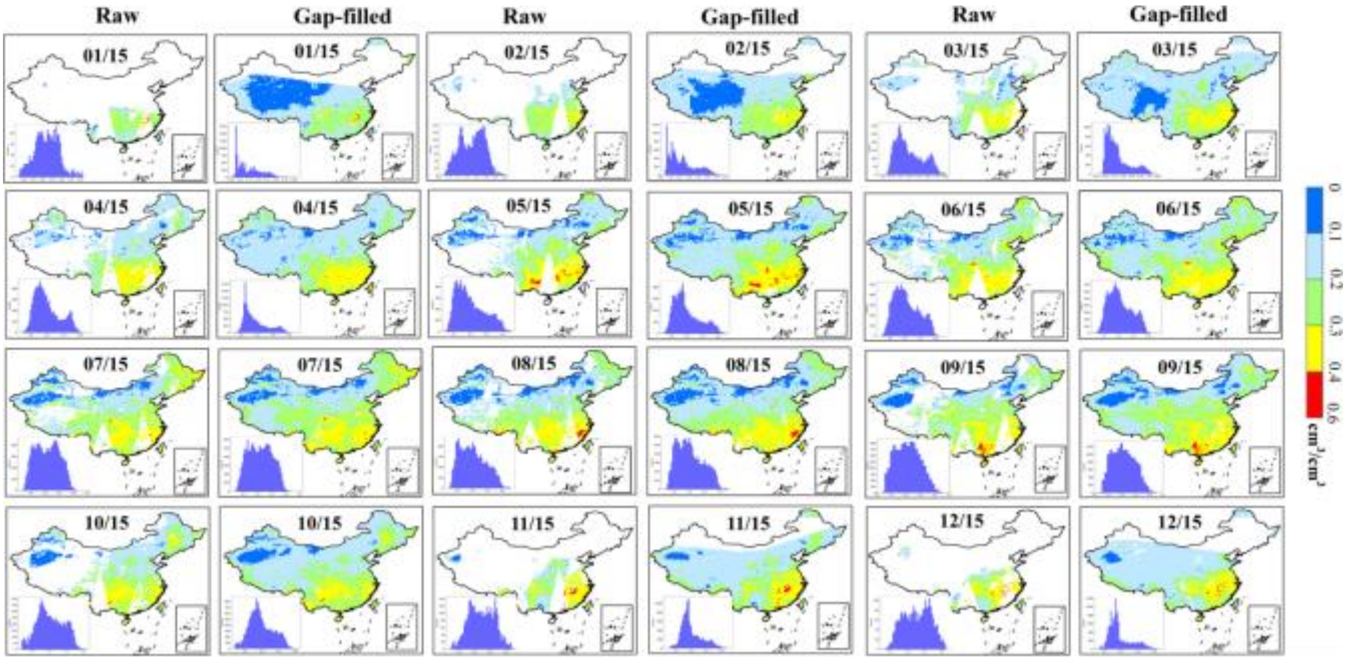


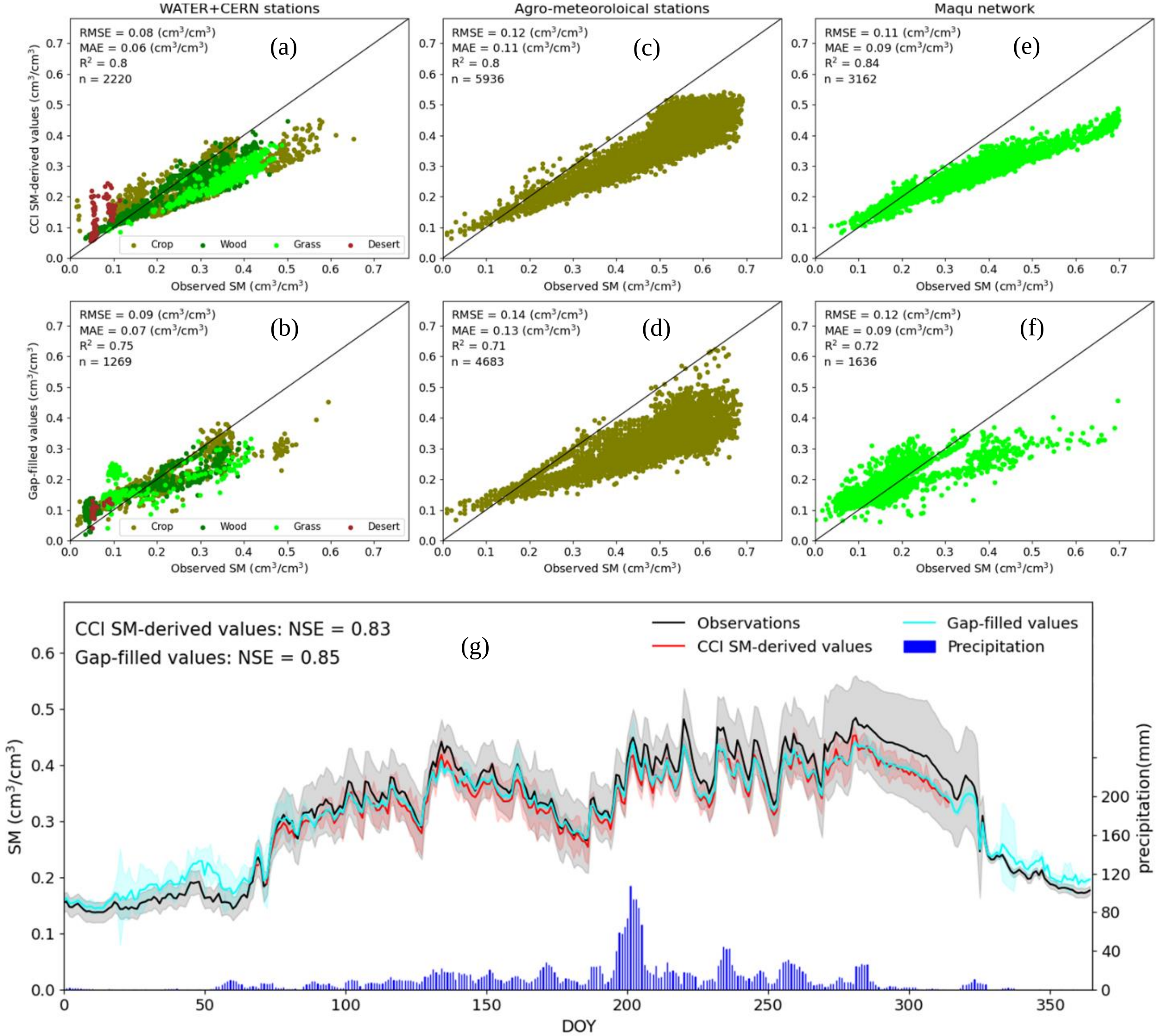
Figure 6: The spatial distributions and histogram of the raw and gap-filled CCI SM on the 15th of each month in 2009.

4.2 Accuracy validation

The proposed model is first evaluated with sparse in situ measurements from WATER and CERN. ~~To avert the mismatch issue, the 25 km SM dataset is disaggregated to 1 km using the DISPATCH model before accuracy validation.~~ As shown in Fig. ~~9~~7(a), ~~good accordance~~ agreement is obtained between the ~~1~~ 1-km CCI SM-derived values and the in situ measurements, with an R^2 of 0.8. This accordance is also found between the ~~1~~ 1-km reconstructed SM and the in situ measurements (Fig. ~~9~~7(b)), with the R^2 of 0.75. ~~Both the CCI SM-derived values and reconstructed SMs are close to the 1:1 line when evaluating with in situ measurements.~~ High ~~accuracies are~~ accuracy is also observed when performing evaluating with in situ measurements from the national agro-meteorological stations. ~~Fig. 9(c) and (d) show the~~The R^2 between the ~~1~~ 1-km CCI SM-derived values and the in situ measurements is 0.81, while the R^2 value between ~~1~~ 1-km reconstructed SM and the in situ measurements is 0.71 (Fig. 7(c) and (d)). ~~In general, our model is capacity in reconstructing SM. The inconsistency still exists,~~ Inconsistency evidently remains, and noticeable overestimations are observed in the high ~~SMs~~ range of SM. Additionally, the accuracy of the gap-filling products tends to be diminished by drought conditions, but this impact is limited.

We further validate the reconstructed results with the dense in situ measurements from the Maqu network. The RMSE and MAE values ~~is~~ was 0.11 and 0.09 cm^3/cm^3 (Fig. ~~9~~7(e)), respectively, for the ~~1~~ 1-km CCI SM-derived values~~, respectively~~, and ~~is~~ 0.12 and 0.09 cm^3/cm^3 (Fig. ~~9~~7(f)), respectively, for the ~~1~~ 1-km reconstructed SM~~, respectively~~. ~~This~~ It means ~~a good~~ that reasonable agreement is obtained for both the CCI SM product and the gap-filled SM~~.~~; however, ~~The~~ poor performance is found in the range of low values~~,~~ mostly ~~due to~~because of the extreme conditions and the fewer samples available for model regression.

The time series of average 0.25^o CCI SM values and reconstructed SM over the dense grid are compared ~~to~~ with the dense in situ observations. Both the original and reconstructed SM ~~matches~~ match well with the in situ series, with ~~the~~ NSE values of 0.83 and 0.85, respectively. The reconstructed SM (Fig. ~~9~~7(g)) mostly describes the temporal dynamics of in situ measurements~~, i.e.~~ ; that is, sufficiently capturing seasonal and daily variability. ~~It is also observed that~~ In addition, the rainfall events impacting the surface dynamics are ~~well delineated on~~ in observed to be well depicted the SM temporal variations. ~~In general, the~~The reconstructed SM ~~seems~~ appears to have inherited the merits of stability between April and November from CCI SM, i.e., having comparable values during this period. ~~This is reasonable, since in addition to focusing on common explanatory variables such as Albedo, NDVI and DEM, our method introduces time series water-heat components, i.e., precipitation, PET, and reanalysis soil moisture, providing a substantial contribution in reconstructing time series SM. Meanwhile, since the used variables are daily, our model shows strong capacity in delineating abrupt climatic changes (Piles et al., 2016).~~

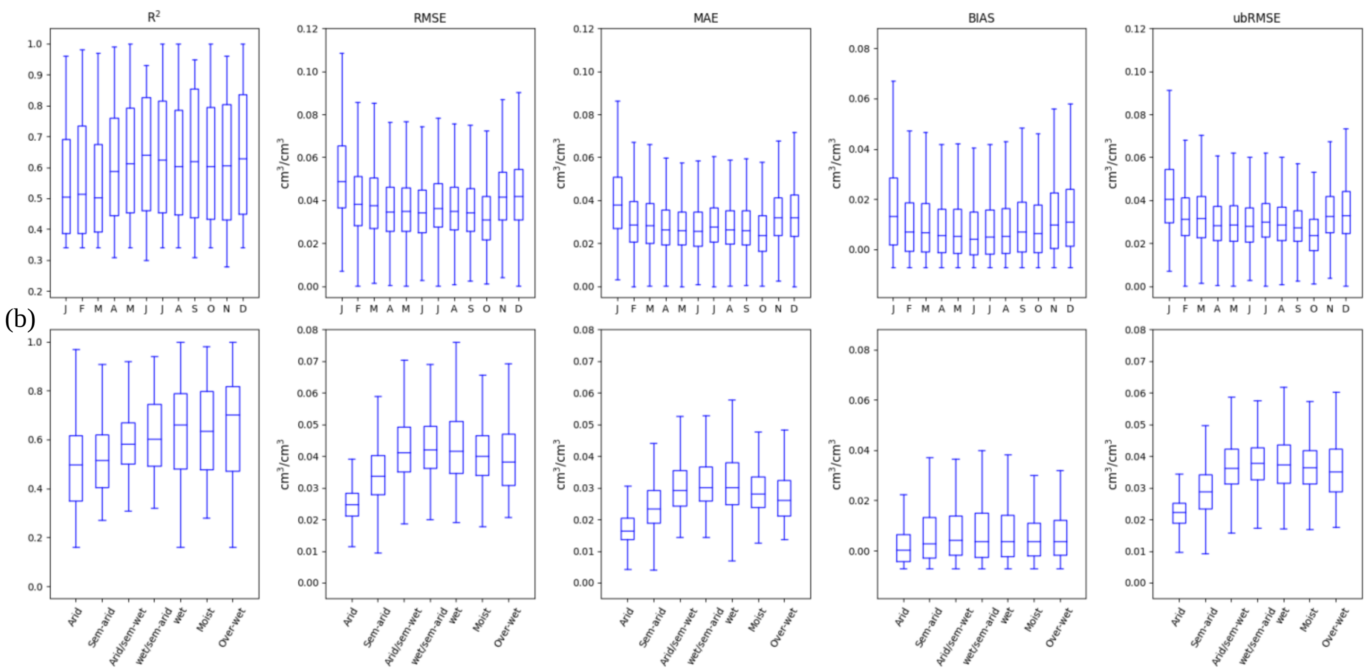


~~Figure 9: The evaluations of model results. (a), (c) and (e) are the scatter plots of 1-km CCI SM-derived values against field measures regarding WATER/CEERN, agro-meteorological stations, and Mauqu network, respectively, and (b), (d) and (f) are the scatter plots of 1-km gap-filled SM-derived values against field measures. (g) are the time series of average CCI SM-derived values against site measures in the Maqu region. The shaded area in (g) denotes ±1 standard error.~~

~~One cross-validation~~ Cross-validation analysis is further ~~conducted~~ performed with 2009 data to evaluate ~~the~~ model performance. The obtained metrics (Fig. ~~10~~8(a)) illustrate ~~a good~~ reasonable coincidence between the reconstructed and the original CCI SM, with ~~the~~ a median R^2 range ~~between~~ of 0.51 and 0.63. Better ~~accuracies are~~ accuracy is also demonstrated ~~in~~ by the metrics of RMSE, MAE and ubRMSE. In particular, the ~~medians~~ median of BIAS ~~are~~ is less than 0.01 cm^3/cm^3. ~~Relatively~~ Comparatively, better accuracies ~~happen~~ is achieved in the growth seasons (March-October)~~. This~~, which can be attributed to the fact that the critical environmental factors, such as NDVI, DTR and ERA soil moisture, are more related to

~~soil moisture~~ satellite-derived SM during the season of vegetation ~~growing seasons~~growth (Chen et al., 2014; Otkin et al., 2016).

Figure ~~10~~8(b) shows the accuracy metrics for different climate regions. A pattern similar ~~pattern as those in~~to that of the monthly means is observed, ~~i.e.~~that is, acceptable accuracy occurs in most regions. No significant differences in median R^2 and BIAS ~~happen~~ are evident between the reconstructed SM of each climate region, with the bias between the maximum and minimum median R^2 and BIAS values being less than 0.09 and 0.003 cm^3/cm^3, respectively. The metrics indicate relatively poor performance ~~of metrics (e.g., RMSE and MAE)~~ in wet regions ~~can be related to the~~having high ~~values in albedo and~~ specific heat capacity and low albedo~~(Guan et al., 2009)~~. ~~Meanwhile, the~~ ~~fewer~~ The lower amounts ~~of~~ and high thermal entropy of the available ~~observations~~ variables (i.e., LST and albedo) in these areas can affect the model capacity and stability (Wang et al., 2005). ~~It should be noticed that~~Notably, despite the relatively high RMSE, MAE and ubRMSE ~~over~~ values in the humid region, the R^2 value is very high~~, as illustrated in~~ (Fig. 10~~.~~), ~~This~~ which might be ~~due~~ attributable to the high SM variability in these areas~~ being high~~. The accuracy is lower over the regions that experience drought due to perturbations of the soil water content, but without noticeably poor performances.



~~**Figure 10: The accuracy metrics of 10-cross validation for R2, RMSE, MAE, BIAS, and ubRMSE: (a) is averagely obtained on a month basis, and (b) is averagely obtained for each climate region.**~~

The spatial distributions of the accuracy metrics in Fig. ~~11~~ 9 further illustrate the ~~good~~ accuracy of the proposed gap-filling model. ~~The obtained metrics in the cross-validation analysis show a good match between the reconstructed SM and the original SM CCI for most of the grids.~~ Discrepancies are observed in some grids, but they rarely exceed 0.09 cm^3/cm^3 in absolute value. Spatially, the distribution of reconstructed SM follows a geographic gradient. The relatively lower accuracies occur ~~on the complicated~~ in areas of complex terrain in western China. For these regions, complex atmospheric conditions caused by high elevations tend to affect the ~~delineations~~ simulation of surface parameters. ~~Meanwhile, complex~~Complex topography ~~may~~ can result in a complicated directional anisotropy, bringing ~~more~~ great uncertainty in ~~modeling~~ modelling surface energy and water cycles (Hu et al., 2016).

The gap-filling model could be sensitive to irrigation and drought owing to the induced inhibition and water stress of vegetation. On the one hand, lower accuracy is found as expected over a considerable fraction of irrigated cropland (e.g., Northern China), which can be partly attributed to the human irrigation drain. On the other hand, focused analyses illustrate the consistency of the gap-filling SM with the in-situ measurements and the original SM under extremely dry conditions (Fig. S4), illustrating the physical plausibility of the gap-filled values for specific application.

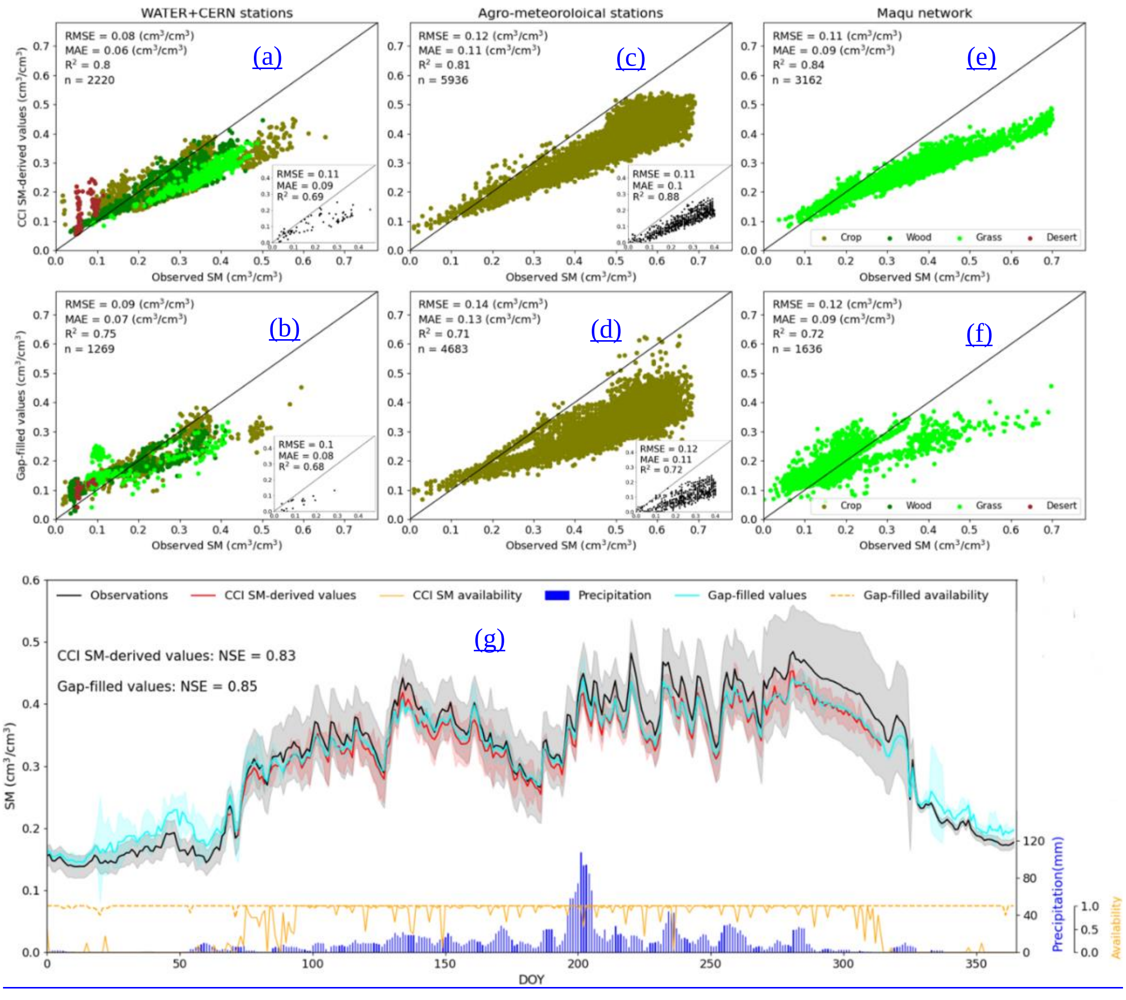


Figure 7: The evaluations of model results. (a), (c) and (e) are the scatter plots of 1-km CCI SM-derived values against field measures regarding WATER/CEERN, agro-meteorological stations, and Mauqu network, respectively, and (b), (d) and (f) are the scatter plots of 1-km gap-filled SM-derived values against field measures. The sub-figures in the upper corner of (a)-(d) are the scatter plots under extremely dry conditions. (g) are the time series of average CCI SM-derived values against site measures in the Maqu region. The shaded area in (g) denotes ±1 standard error.

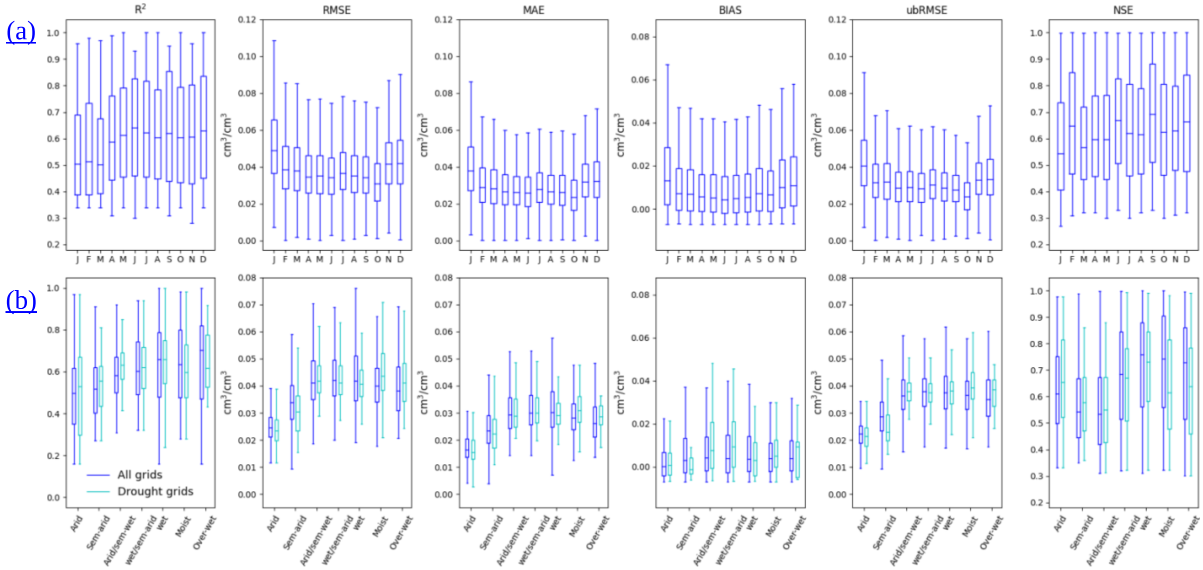
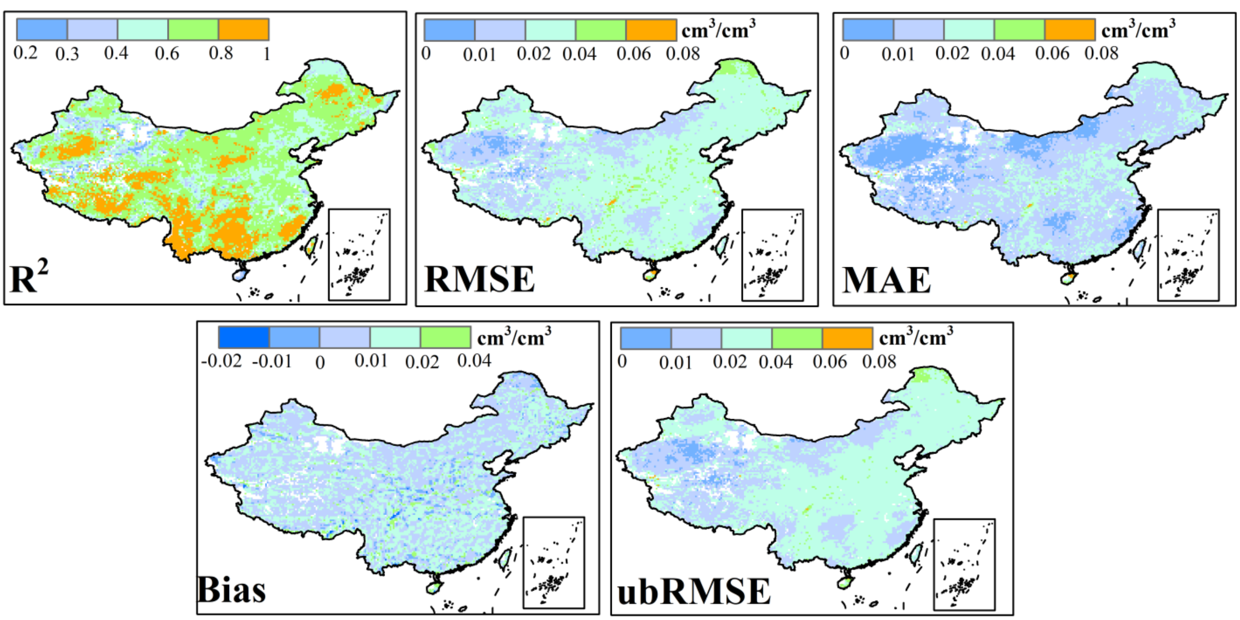


Figure 8: The accuracy metrics of 10-cross validation for R^2, RMSE, MAE, BIAS, ubRMSE and NSE: (a) is averagely obtained on a month basis, and (b) is averagely obtained for each climate region and for the drought grids.



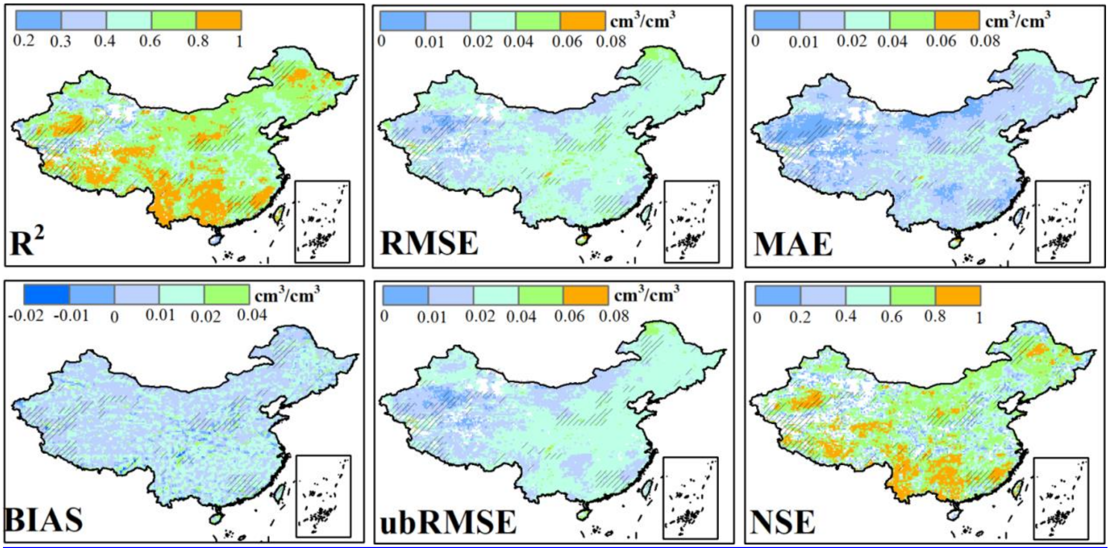


Figure ~~11~~9: The spatial distributions of accuracy metrics of 10-cross validation in 2009 for R2, RMSE, MAE, BIAS, and ubRMSE and NSE. The slash represents the regions impacted by drought.

4.3 Comparison analysis

~~To fully reveal the merits in modeling the critical surface characteristics, our model is further compared against previous studies, as illustrated in Table 4. In general, the accuracies produced by our model are comparable and even better than previous studies, despite being evaluated with different in suit measurements and simulated approaches. The satisfied performance of the proposed model is plausible, which can be attributed to the following aspects. Given the complex underlying surface and the diverse climatic conditions across one continuous scale, we chose more efficient explanatory variables relative to previous research. Especially, the land surface model and reanalysis outputs are introduced to bring additional context for overcoming the severe missing issues of remotely sensed variables. Simultaneously, an adaptive spatiotemporal domain strategy and residual calibration module are incorporated into machine learning to balance the regression performance, overfitting problems, and computational complexity. This strategy focuses on the covariates that include both the spatial and temporal domains, therefore possessing the potential of producing more reasonable accuracy compared to other approaches that utilize either spatial domain (Li et al., 2021a).~~

~~Table 4 The accuracy comparison among different literatures~~

	~~SM source~~	~~Model~~	~~Adopted variables~~	~~Study Region~~	~~Accuracy~~		~~Literatures~~
					~~Field validation~~	~~simulation validation~~	
~~1~~	~~ESA CCI~~	~~RF~~	~~LST, Precipitation, NDVI, PET, Soil texture~~	~~Southern Europe~~		~~RMSE=0.024-0.025 m^3/m^3~~	~~Almendra-Martín et al. (2021)~~

~~2~~	~~ESA CCI~~	~~ANN~~	~~Precipitation, Temperature, NDVI, LAI, DEM, Slope, Aspect, Latitude, Longitude, Soil texture~~	~~China~~	~~RMSE=0.036-0.074 m^3/m^3~~	~~RMSE=0.037-0.064 m^3/m^3~~	~~Zhang et al. (2021b)~~
~~3~~	~~ESA CCI~~	~~RF~~	~~NDVI, LST, Daytime LST, Nighttime LST, Diurnal LST~~	~~Oklahoma, USA~~	~~RMSE=0.08 m^3/m^3~~	~~RMSE=0.02 m^3/m^3~~	~~Liu et al. (2020b)~~
~~4~~	~~AMSR2~~	~~CNN~~	~~/~~	~~Global~~	~~RMSE=0.097 m^3/m^3~~	~~RMSE=0.065-0.073 m^3/m^3~~	~~Zhang et al. (2021c)~~
~~5~~	~~ESA CCI~~	~~MLR, OK, RK~~	~~Precipitation, Temperature~~	~~Midwest, USA~~	~~RMSE=0.067-0.070 m^3/m^3~~		~~Llamas et al. (2020)~~
~~6~~	~~FY-3B~~	~~BP-NN~~	~~LST, NDVI, Albedo, Latitude, Longitude, DEM, DOY~~	~~Tibetan Plateau~~	~~RMSE=0.1 cm^3/cm^3~~		~~Cui et al. (2016)~~
~~7~~	~~ESA CCI~~	~~RF~~	~~Albedo, NDVI, DTR, AP, PET, ERA SM and TPI~~	~~China~~	~~RMSE=0.09-0.14 cm^3/cm^3~~	~~RMSE=0.025 cm^3/cm^3~~	~~Our study~~

~~Note: CNN = Convolutional neural network, OK = ordinary kriging, RK = regression kriging, BP-NN = back-propagation neural network~~

The proposed method is future compared against four extensively used models, and the accuracy metrics of the five models are shown in Fig. 10. ~~that adopt the same explanatory variables and spatiotemporal window search strategy. The first one is the conventional multiple linear regression (MLR) approach. Three typical machine learning approaches, i.e., Extreme gradient boost (XGB), Support vector machine (SVM) and Artificial Neural Network (ANN), are also used for comparison. Detailed descriptions of four available models can be found in supplementary Text S2. The accuracy metrics of the five models are shown in Fig. 12.~~ Generally, the MLR, XGB, SVM and ANN, accompanying the RF, could potentially reconstruct the missing CCI SM pixels, indicating the stable suitability of these models and the feasibility of available variables. Moreover, the RF model demonstrates prominent performance among all the tested models, further ~~manifesting~~ demonstrating its capacity ~~in~~ for reconstructing ~~soil moisture~~SM when integrating ~~the~~ an effective dataset source and mining method. ~~This is~~ Our results are consistent with earlier studies that ~~illustrate~~ illustrated the robustness of the RF approach in simulating satellite parameters (Karbalaye Ghorbanpour et al., 2021; Zhao et al., 2018). This is attributed to the strong capacity of the RF method for coping with sparse samples, in addition to the fact that the RF does not assume a specific functional or geometric form of the model. ~~Based on this, we~~We also check the accuracy of the models excluding the residual calibration procedure, which is ~~one~~ an essential ~~procedure~~ component for the proposed model. Results~~ in Fig. 12~~ (in Fig. 10) demonstrate the accuracies are lowered by ~9% when removing residual calibration, underscoring the importance of residual modulation in improving SM reconstructing.~~ Meanwhile~~ Moreover, ~~the~~ better performance brought by the spatiotemporal domain strategy is also exhibited when compared with the global regression~~, as illustrated in Fig. 12~~. Quantitatively, the spatiotemporal domains can improve the accuracy of ~19% in forcing the RF regression. Overall, these analyses indicate the feasibility of the proposed model by integrating the modules of residual calibration and spatiotemporal domain strategy.

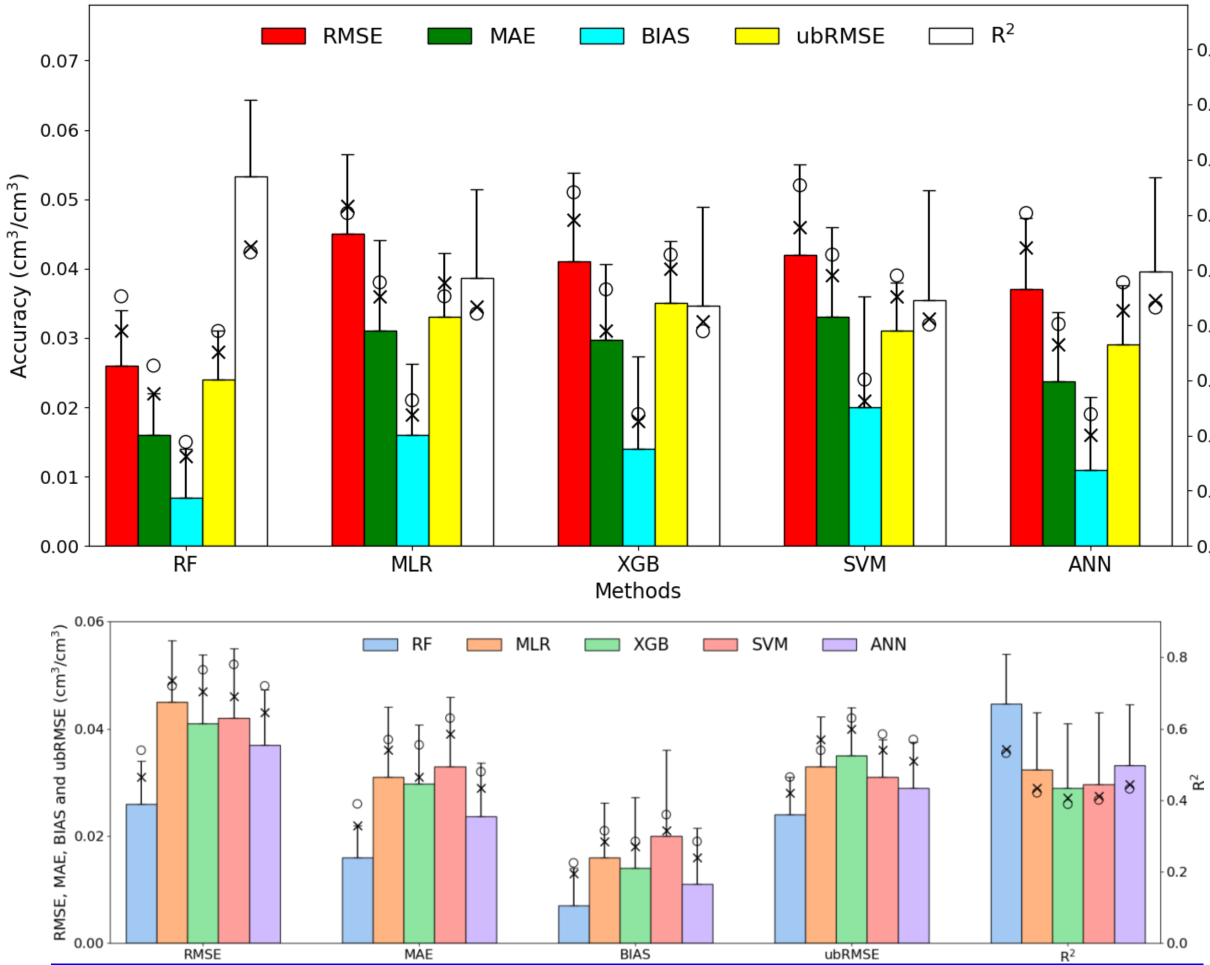
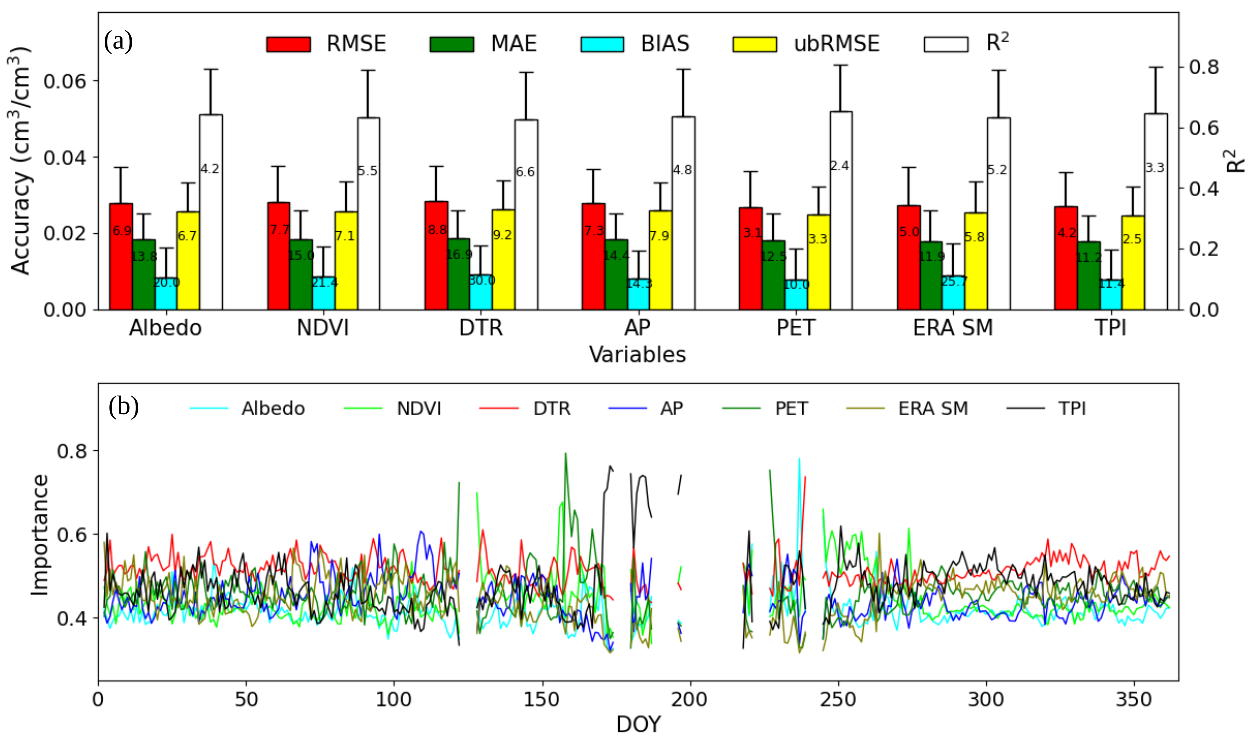


Figure ~~12~~10: Comparison RF-based model with other models (i.e., MLR, XGB, SVM and ANN). Error bars denote 1σ errors. The symbol 'x' represents the accuracy metrics of models excluding the residual calibration procedure, and the symbol 'o' represents the accuracy metrics of the models that use the global regression rather than regional regression based on the spatiotemporal window searching strategy.

4.4 Importance analysis

~~Considering the criticality of explanatory variables in simulating SM, importance analyses regarding these selected variables are conducted.~~ We ~~first~~ investigate the accuracy of the reconstruction model that excludes one participating variable. As illustrated in Fig. ~~13~~11(a), the performance of the model with six variables (i.e., excluding one) is relatively ~~lower~~ low when compared with that of a model with seven variables. The ~~removing one~~ strategy of removing one variable can lower the accuracies of 2.2-6.4% in terms of R^2 and by 10-30% in terms of BIAS. This diminished performance is plausible ~~since~~ because SM is heavily related to all the selected variables. Specifically, variability in land surface characteristics (NDVI and albedo) and atmospheric conditions (i.e., precipitation and PET) can ~~heavily~~ impact ~~the soil moisture~~SM variability. This is plausible because satellite SM retrievals represent the signals from the upper soil layer, which is directly exposed to the land and the atmosphere. Meanwhile, additional covariates mean an increase in the number of samples participating in the regression model, therefore potentially resulting in ~~an~~ improvement of overall accuracy. ~~Specifically, we~~We observe that the lowest accuracy occurs when DTR is excluded, underscoring the vital role of DTR in ~~modeling~~ modelling SM.

~~We further investigate the contribution of each explanatory variable in modeling SM.~~ The importance ~~score~~ scores produced by the RF algorithm (Zhao et al., 2019b; Ramoelo et al., 2015) (Fig. S5) also show that all selected variables substantially impact the CCI SM simulations~~itself is used to delineate relative contribution (Zhao et al., 2019b; Ramoelo et al., 2015)~~. ~~As demonstrated in Fig. 13(b) (and Fig. S4), AP and ERA SM substantially impact CCI SM modeling, with the average importance score of 0.48 and 0.47, respectively. The NDVI, albedo, TPI and PET have less importance with the average score of 0.41, 0.43, 0.42 and 0.45, respectively.~~ Specifically, ~~the~~ DTR shows the ~~highest~~ greatest importance, ~~with an average score of 0.52. This is~~ mainly ~~related~~ relating to the fact that temperature variations ~~likely have a far-reaching~~might influence ~~on soil moisture~~SM fluctuation. This supports the higher model performance observed in warm seasons, during which PET, albedo, and NDVI exhibit a higher importance score. During this period, ~~The~~ heat from the surface can be transferred to the atmosphere via ~~evapotranspiration~~ ET and sensible heat conduction, thereby modifying surface ~~soil moisture~~SM variations (Amani et al., 2017).



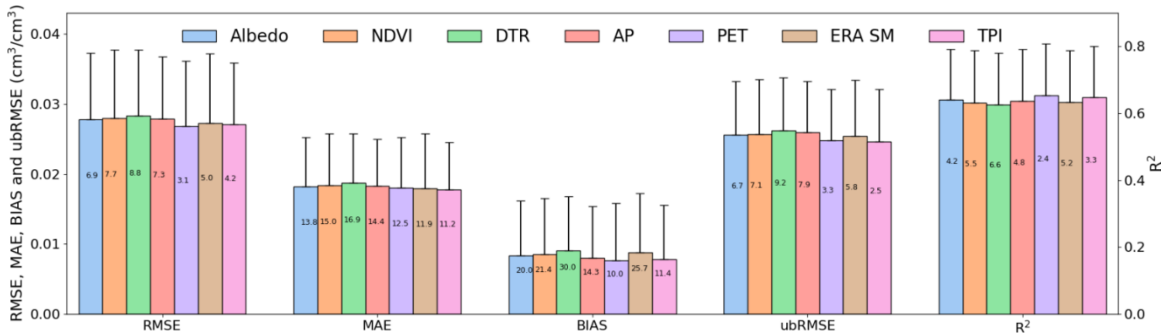


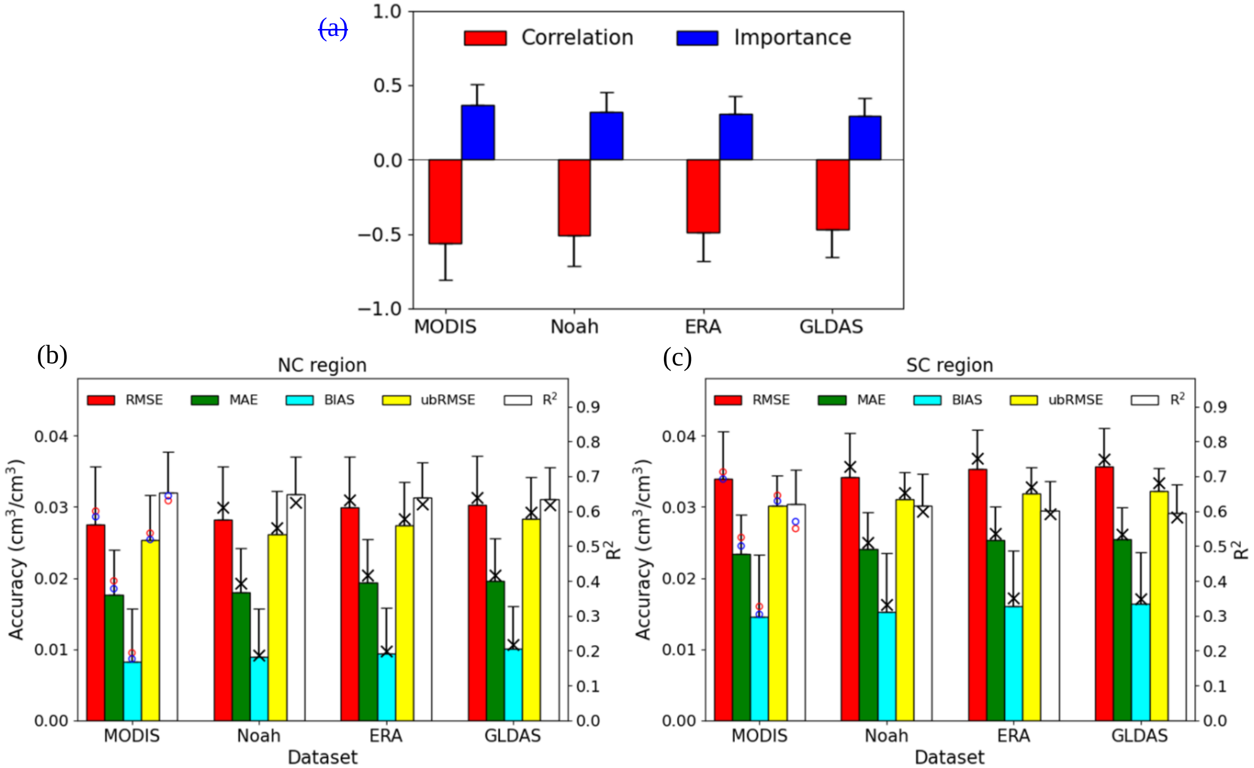
Figure ~~13~~11: ~~(a)~~ The accuracy of the models removing one variable, i.e., using other six variables in model regression. Error bars denote 1σ errors. The text denotes the relative percentage of the ~~lowered~~ decreased accuracy of the model with six variables (i.e., excluding one) in comparison ~~to those using all~~ with that of a model with seven variables. ~~(b) The importance score of the selected variables derived from RF regression.~~

~~**4.5 Uncertainty analysis**~~

~~Given the critical importance of satellite-derived DTR and the severe missing issues in satellite-observed LST products, we~~We further investigate the substitution performance of other surface temperature ~~sources~~ products in reconstructing SM~~,~~ ~~i.e., i.e., Noah, ERA and GLDAS~~. ~~This analysis is conducted at two focused regions (in Fig. 1) that have sufficient data sources supporting our experiments: one region is in northern China that is mostly occupied by arid and semi-arid areas, while the other region is in southern China that is occupied by wet areas.~~ Considering the bias between satellite-derived LST and ~~modeled~~ modelled surface temperature, the variable correction described in section ~~3.1.2~~3.1.5 is conducted to remove the systematic bias and make the simulated DTR comparable ~~to~~ with the satellite observations.

~~Figure 14(a) shows the minor~~ Minor reductions are found in the Pearson correlation and RF-derived importance score of three numerical ~~model~~ model-simulated DTRs (Fig. S6) when ~~comparing~~ compared with the MODIS-derived DTR, which indicates the feasibility of using each of ~~all~~ these datasets in reconstructing SM. ~~Fig. 14(b) and (c) illustrate noticeable reductions~~ Reductions in model accuracy are evident when replacing the satellite-derived LST with the other three simulated sources (Fig. 12(a) and (b)). ~~Regarding the accuracy metric, the RMSE of reconstructed SM can be reached to 0.0282, 0.0299 and 0.0303 cm^3/cm^3, when respectively using Noah, ERA and GLDAS-simulated DTR over the northern region. As for southern China, the RMSE of reconstructed SM can be reached 0.0342, 0.0353 and 0.0357 cm^3/cm^3, when respectively using the other three dataset.~~ Nevertheless, the availability of reconstructed SM products is remarkably increased (by ~6-11%) ~~due~~ owing to the all-weather coverage of the reanalysis and land surface model simulations. ~~This suggests the~~The surface temperature source from the numerical model dataset ~~can be one~~ is suggested to be an alternative for satellite LST, which is essential ~~at one~~on the long-term and large extended scale, especially considering their ~~full~~ full-coverage ~~attribution~~ characteristic. ~~On the other hand, as illustrated in Fig. 14 (b) and (c), one noticeable~~ However, in comparison with the results obtained using the correction procedure, reduction in accuracy metric (~4%) occurs when not considering the variable correction procedure. ~~This~~ It emphasizes the indispensable contribution of the variable calibration procedures in reconstructing surface characteristics (Duan and Bastiaanssen, 2013; Liu et al., 2020a).

~~Since the reanalysis SM is one vital input in our approach, we~~We also compare ~~it~~ the ERA SM with ~~the other~~ two other products to evaluate ~~the~~ its feasibility ~~of ERA data~~ in reconstructing CCI SM. GLEAM and Noah surface SM are respectively employed to replace the ERA SM while keeping other explanatory variables ~~keep the rest~~ the same. Although the GLEAM and Noah SM-based schemes can demonstrate acceptable accuracies, they exhibit slightly inferior accuracies ~~compared to~~ in comparison with ERA SM-based schemes ~~(Fig.14 (b) and (c))~~, probably ~~due~~ owing to their relatively large uncertainties in ~~delineating~~ depicting the surface ~~soil moisture~~SM dynamics across the two selected regions ~~(Mahto and Mishra, 2019; Chen and Yuan, 2020)~~. Nevertheless, ~~considering~~ our study ~~merely~~ focuses on only two local regions~~,~~ ; therefore we cannot ~~manifest~~ claim that the ERA product could provide the best performance across China~~. More~~, and more attention should be focused on this in further work.



~~**Figure 14: The performance of models using Noah, ERA, and GLDAS DTR replacing MODIS DTR. (a) The Pearson correlation and importance score of DTRs. (b) and (c) The metrics of models using different DTRs for Northern China (NC) and Southern China (SC), respectively. Error bars denote 1σ errors. The symbol 'x' represents the accuracy metrics of the models without DTR correction procedure. The symbol 'o' in red represents the accuracy metrics of the models using GLEAM SM to replace ERA SM, and the symbol 'o' in blue represents the accuracy of the models using Noah SM to replace ERA SM.**~~

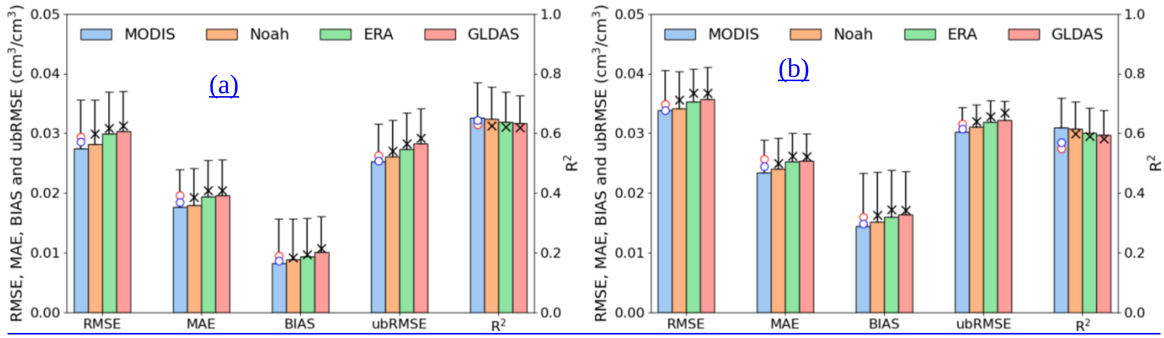


Figure 12: The metrics of models using different DTRs for (a) Northern China (NC) and (b) Southern China (SC). Error bars denote 1σ errors. The symbol 'x' represents the accuracy metrics of the models without DTR correction procedure. The symbol 'o' in red represents the accuracy metrics of the models using GLEAM SM to replace ERA SM, and the symbol 'o' in blue represents the accuracy of the models using Noah SM to replace ERA SM.

~~4.6 Extending to one long-term scale~~4.5 Long-term extension

~~The available dataset forcing our model has a long sequence, implying one potential in modeling long-term SM products. To verify this, the~~The proposed gap-filling method is further extended to the long-term ECA CCI SM databases. During 2005—2015, more than 90% of ~~the~~ contaminated pixels can be reconstructed using our model~~, as illustrated in Fig. 15(a) and (b)~~. When evaluating the pixels against ~~dense~~ in situ measurements ~~in~~ from the dense Maqu network, we observe that the reconstructed SM during 2005—2015 has ~~comparable~~ accuracy ~~with the original SM~~ that is comparable to that in 2009 (Table ~~5~~4). The average R^2 and RMSE values of the reconstructed SM are 0.73 and 0.12 cm^3/cm^3, respectively. The present results indicate the proposed model has ~~a~~ strong capacity ~~to simulate~~for simulation of SM ~~at a~~on the long-time scale.

~~Figure 15(c) shows~~ The spatial distribution and the obvious differences between the gap-filled and original SM dataset can be seen in Figure 13 (a)-(c). Negative differences in SM occur in most regions, while positive differences ~~happen~~ are evident in ~~a~~ small ~~fraction~~ areas of the wet and arid regions. The ~~dynamic~~ dynamics and ~~trend~~ trends of SM are fundamental to assessing and quantifying ~~eco-hydrology~~ ecohydrological regime. Owing to the missing satellite retrievals, the CCI SM tends to be overestimated. ~~Accordingly, we further investigate the trend of SM series during 2005-2015, which is obtained with Sen's slope and M-K significant analysis (Li et al., 2021b; Liu et al., 2021a).~~ As shown in Fig. ~~15~~13(d)-(f), the difference in valid participating SM values ~~participating in trend analysis brings about a noticeable disparity in SM trend, which implies a slightly decreased SM trend for most arid regions in China and an increased SM trend for most humid regions~~ causes disparity in calculating the SM trend, i.e., bringing a lower SM trend in most wet regions but a higher SM trend in some dry regions when gap-filled values are introduced. Additionally, most regions with a significant trend demonstrate a lower trend in comparison with the trends of the original SM. The confidence level of the SM trend is converted from a significance level to a non-significance level for a considerable fraction of the grids. This is more pronounced in wet regions such as northeast, northwest, and southwest parts of China, which are sensitive to monsoon precipitation and ice melting. Our results are corroborated by earlier studies (Zhang et al., 2018; Gunnarsson et al., 2021) that revealed an overestimation in the trend of missing AOD and albedo when cloudy conditions prevented satellite retrievals. ~~This pattern~~ It means that the variations in SM trend ~~is mostly~~ are related to ~~the climate~~ changes in the climate variables (e.g., precipitation ~~and temperature~~) ~~changes and human~~ and land management activities (Li et al., 2018).

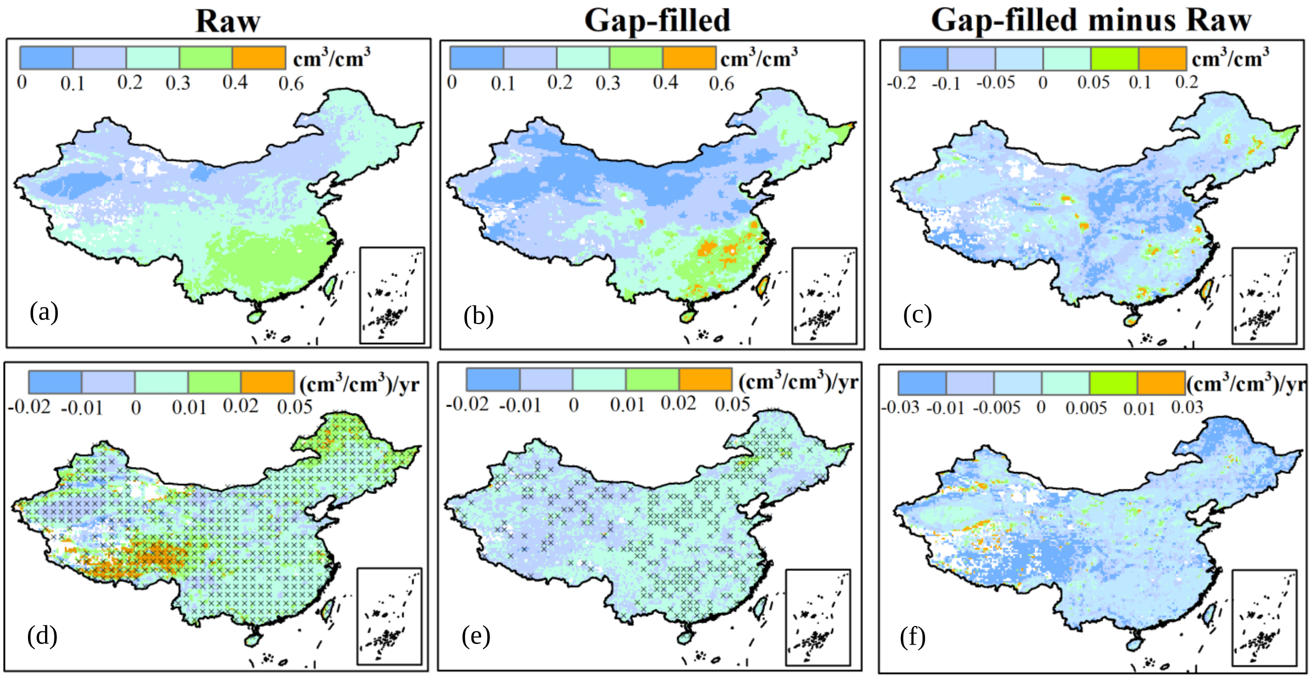
~~The biases in SM dynamic and trend are more pronouncedly delineated for each climate region in Fig. 16(a) and (b). Results show that the trends from the reconstructed SM are relatively lower compared to those from the original CCI SM. The improvement of the reconstructed dataset in depicting SM trends are quantitatively manifested in Fig. 16(c) and (d), which demonstrates the R^2 between the trends from the original CCI SM and those from in situ measurement is 0.23 while the R^2 between the trends from the reconstructed CCI SM and those from observations is increased to 0.45. Overall, an effective gap-filled model is demanded considering its capacity in fully depicting dynamics and trends of SM.~~

Table ~~5~~ 4 Metrics for the gap-filling performance regarding Maqu network for the extended years

Year	R^2		RMSE (cm^3/cm^3)		MAE (cm^3/cm^3)		Bias (cm^3/cm^3)		ubRMSE (cm^3/cm^3)		NSE	
	CCI	gap-filled	CCI	gap-filled	CCI	gap-filled	CCI	gap-filled	CCI	gap-filled	CCI	gap-filled
2008	0.8	0.71	0.11	0.13	0.1	0.13	0.06	0.07	0.06	0.06	0.8	0.81
2010	0.82	0.73	0.1	0.11	0.09	0.11	0.05	0.06	0.06	0.05	0.81	0.83
2011	0.83	0.74	0.09	0.11	0.09	0.1	0.06	0.06	0.06	0.05	0.82	0.84
2012	0.81	0.72	0.12	0.13	0.09	0.12	0.06	0.05	0.05	0.05	0.81	0.82
2013	0.82	0.73	0.09	0.12	0.09	0.13	0.06	0.07	0.05	0.07	0.8	0.82
2014	0.85	0.74	0.09	0.11	0.08	0.09	0.06	0.08	0.05	0.06	0.83	0.85
2015	0.79	0.69	0.12	0.14	0.1	0.12	0.07	0.09	0.07	0.07	0.79	0.81

~~Year~~	~~Metrics for raw CCI dataset~~						~~Metrics for gap-filled dataset~~					
	~~R^2~~	~~RMSE~~	~~MAE~~	~~Bias~~	~~ubRMSE~~	~~NSE~~	~~R^2~~	~~RMSE~~	~~MAE~~	~~Bias~~	~~ubRMSE~~	~~NSE~~
~~2008~~	~~0.8~~	~~0.11~~	~~0.1~~	~~0.06~~	~~0.06~~	~~0.8~~	~~0.71~~	~~0.13~~	~~0.13~~	~~0.07~~	~~0.06~~	~~0.81~~
~~2010~~	~~0.82~~	~~0.1~~	~~0.09~~	~~0.05~~	~~0.06~~	~~0.81~~	~~0.73~~	~~0.11~~	~~0.11~~	~~0.06~~	~~0.05~~	~~0.83~~
~~2011~~	~~0.83~~	~~0.09~~	~~0.09~~	~~0.06~~	~~0.06~~	~~0.82~~	~~0.74~~	~~0.11~~	~~0.1~~	~~0.06~~	~~0.05~~	~~0.84~~
~~2012~~	~~0.81~~	~~0.12~~	~~0.09~~	~~0.06~~	~~0.05~~	~~0.81~~	~~0.72~~	~~0.13~~	~~0.12~~	~~0.05~~	~~0.05~~	~~0.82~~
~~2013~~	~~0.82~~	~~0.09~~	~~0.09~~	~~0.06~~	~~0.05~~	~~0.8~~	~~0.73~~	~~0.12~~	~~0.13~~	~~0.07~~	~~0.07~~	~~0.82~~
~~2014~~	~~0.85~~	~~0.09~~	~~0.08~~	~~0.06~~	~~0.05~~	~~0.83~~	~~0.74~~	~~0.11~~	~~0.09~~	~~0.08~~	~~0.06~~	~~0.85~~
~~2015~~	~~0.79~~	~~0.12~~	~~0.1~~	~~0.07~~	~~0.07~~	~~0.79~~	~~0.69~~	~~0.14~~	~~0.12~~	~~0.09~~	~~0.07~~	~~0.81~~

Note: NSE is from the evaluation with the time series of average 0.25^o pixels while the other five metrics are from the evaluation with 1 km disaggregated values.



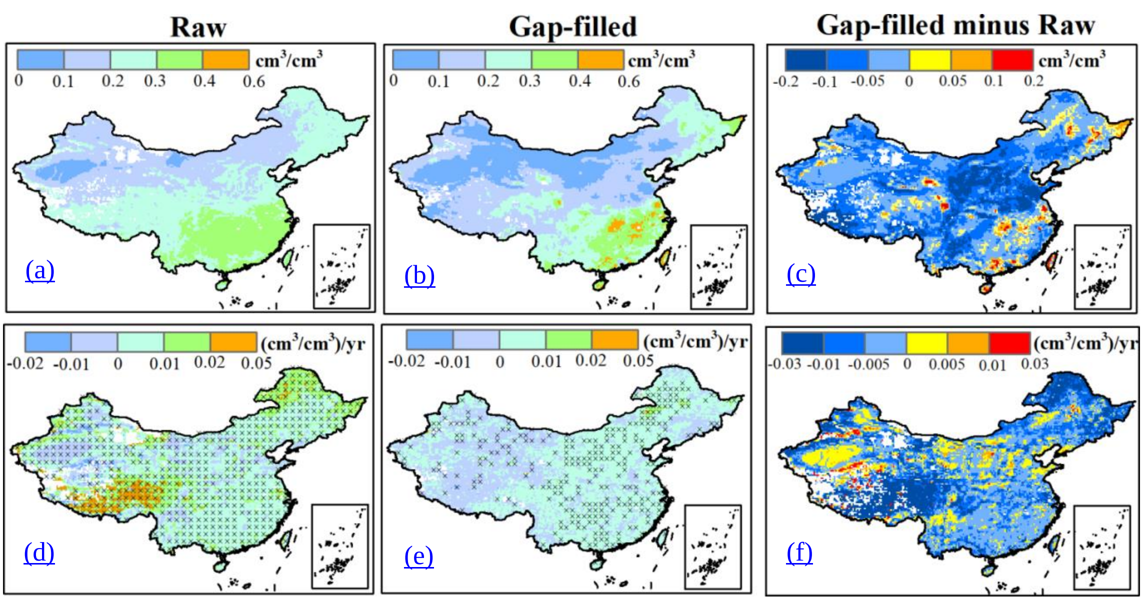
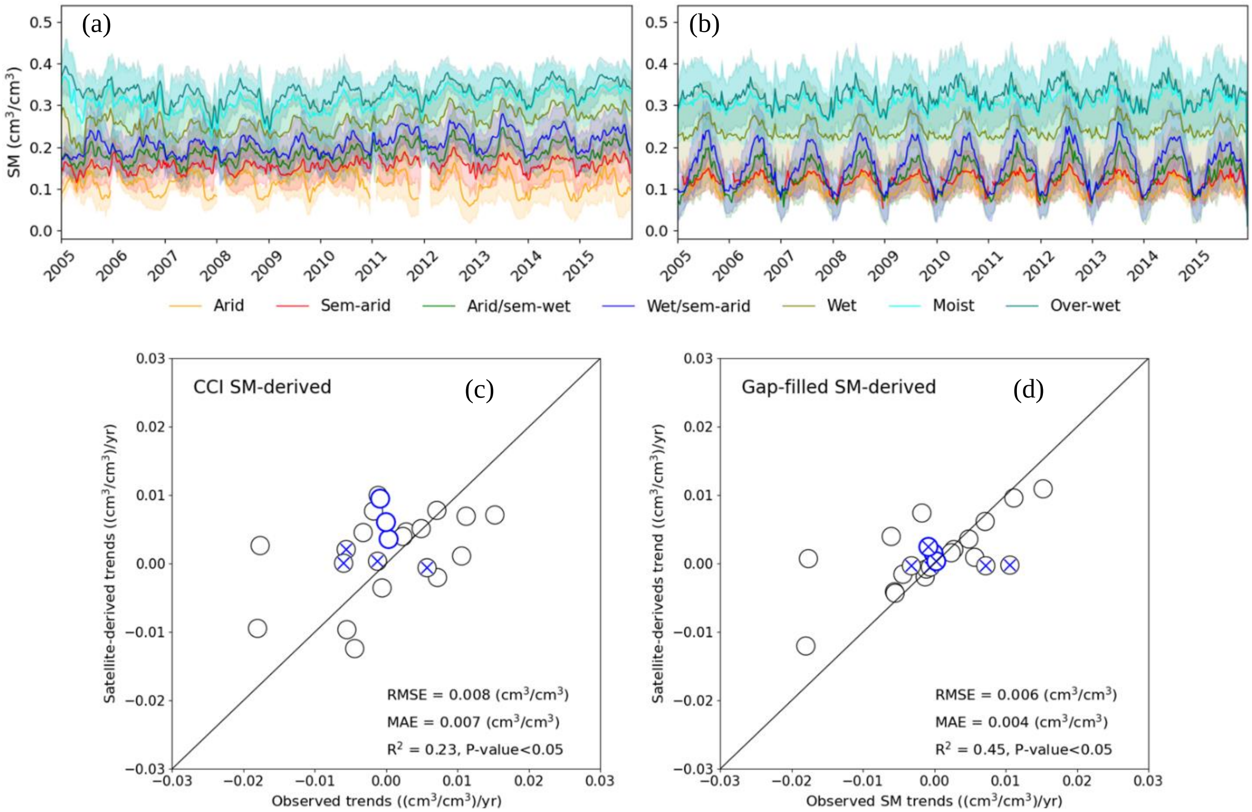


Figure ~~15~~13: The implementation of the proposed model to 2005-2015. (a) and (b) are the average values of raw CCI and gap-filled SM during 2005-2015, and (c) is the difference between them. (d) and (e) are the average trends of raw CCI and gap-filled SM during 2005-2015, and (f) is the difference between them. The symbol "x" in (d) and (e) denotes the significance level under 0.05.



~~**Figure 16: (a) shows the temporal patterns of raw CCI SM regarding different climate regions during 2005-2015, and (b) shows the temporal patterns of gap-filled CCI SM regarding different climate regions. The shaded area in (a) and (b) denotes ±1 standard error. (c) The scatter plot of 1-km CCI SM-derived trends against in situ measures during 2005-2014, and (d) the scatter plot of 1-km gap-filled SM-derived trends against in situ measures. The blue circle in (c) and (d) means the trends from in situ measures are under insignificance level, while the fork means the trends from satellite-derived SM are under insignificance level.**~~

The biases in SM dynamics and trends are shown more pronounced for each climate region in Fig. 14(a) and 14(b). The regional averages of reconstructed SM are relatively low in comparison with those from the original CCI SM, and this pattern is clearly reflected in the trend-cycle and seasonal component (Fig. S7). The improvement of the reconstructed dataset in depicting SM trends is quantitatively manifested in Fig. 14(c)-(f), that is, the R2 value between the trends from the original CCI SM and those from the in situ measurements is 0.28, while the R2 value between the trends from the reconstructed CCI SM and those from the observations is increased to 0.49. Overall, an effective gap-filled model is demanded considering its capacity from depicting the dynamics and trends of SM.

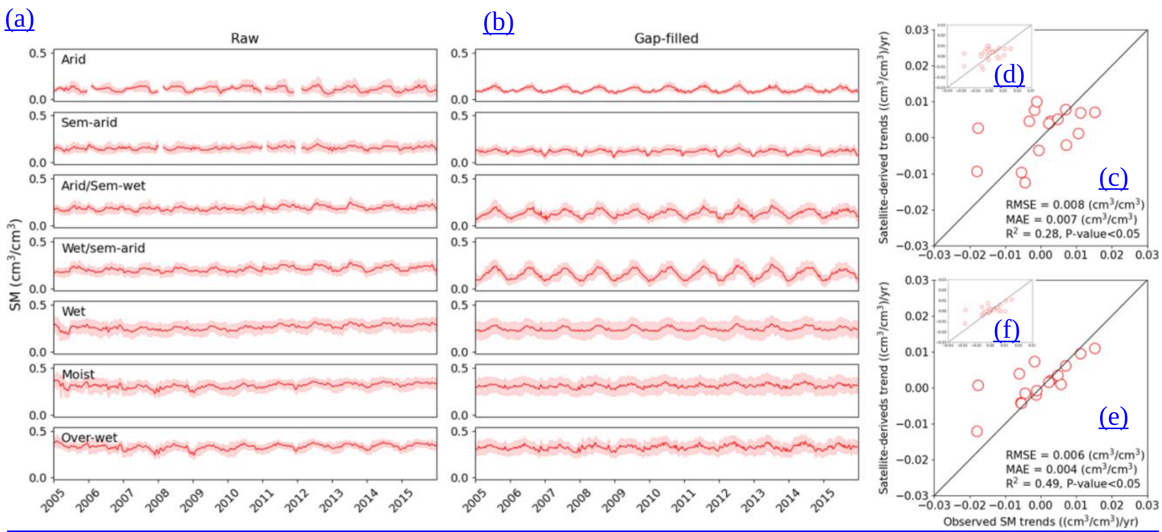


Figure 14: (a) shows the temporal patterns of raw CCI SM regarding different climate regions during 2005-2015, and (b) shows the temporal patterns of gap-filled CCI SM regarding different climate regions. The shaded area in (a) and (b) denotes ±1 standard error. (c) and (d) The scatter plot of 1-km CCI SM-derived trends against in situ measures during 2005-2014, and (c) shows the trends under significance level, while (d) shows all the trends. (e) and (f) The scatter plot of 1-km gap-filled SM-derived trends against in situ measures during 2005-2014, and (e) shows the trends under significance level, while (f) shows all the trends.

5. Conclusions and future considerations

The continuity of satellite-derived SM series is hampered by ~~the~~ data gap ~~issues~~problems. This study thus provides a novel framework for reconstructing a spatially continuous daily SM dataset by integrating the ~~ESA~~ European Space Agency CCI SM and ~~the~~ related explanatory variables. To achieve this, ~~one~~ the random forest method taking full account of both the spatial and temporal domains is ~~designed~~ adopted. The explanatory variables filtered based on ~~the basis of~~ a spatiotemporal window search strategy exhibit ~~a~~ substantial effect in driving the RF regression, resulting in ~~an~~ efficacy improvement of ~19%. Meanwhile, model performance is enhanced by calibrating the derived residuals based on ~~GWR~~ geographically weight regression and Gaussian filters. This improvement is manifested by the fact that the accuracies of gap-filling models are lowered by ~9% when removing the residual calibration procedure.

~~Study presents~~ Our study illustrates the merit of identifying a sufficient number of explanatory variables from the integration of satellite observations and model-driven knowledge. This is clearly verified by the fact that the accuracy of reconstructed SM is noticeably reduced when excluding one of each of the participating variables in turn while retaining the remaining variables. The selected variables complementarily reproduce the SM dynamics in addition to capturing the spatial variations, which also implies that the nonlinear correlation between the SM and explanatory variables can be ~~delineated~~ depicted on the spatiotemporal scale. ~~Importance analysis illustrates that the accuracy of reconstructed SM is noticeably reduced when excluding each participating variable while keeping the rest variables. Specifically, in~~In addition to conventional variables from optical remote sensing, the essential environment elements from model-driven knowledge are used to improve the performance of SM reconstruction. Earlier studies have suggested (Li et al., 2021a; Long et al., 2019) that ~~the~~ reanalysis dataset and land surface model ~~product can~~ products could provide spatiotemporally continuous records, indicating the great potential of simulating land surface parameters. Our study proposes to merge CCI SM time series with the reanalysis and land surface model dataset and applies to China. Here, we employ a machine learning model and a bias correction procedure for CCI SM simulation, which is expected to leverage the knowledge of the reanalysis dataset and the output from the land surface model in transfer to the CCI SM time series. The reconstructed SM achieves ~~satisfying~~ satisfactory accuracy over China~~, especially for areas with large swath gaps~~, underscoring the importance of spatial coverage and continuity of the environmental factors from model-driven knowledge, and ~~the~~ highlighting the need for multiple datasets ~~should~~ to be involved in ~~the~~ gap-filled models. We further confirm this with ~~one~~ an uncertainty analysis showing the feasibility of using alternative data sources of DTR and SM, which is essential ~~at one~~on the long-term ~~scale~~scales, ~~especially~~ considering the full coverage ~~attribution~~ characteristic of numerical model simulated products. Nevertheless, ~~since the~~ because numerical simulation ~~model simulated~~ models are generally sensitive to regional surface and climatic conditions, adoption of more effective machine learning models and bias correction strategies, as well as more representative model ~~products~~ outputs such as CLDAS and regional numerical models, ~~can~~ could be considered in further work (Li et al., 2022a; Li et al., 2022b).

Machine learning is ~~previously reported to be~~ recognized as a powerful ~~strategy~~ tool for reconstructing contaminated values. Despite the effectiveness of the RF ~~models~~ model in situ SM databases, its applicability in reconstructing long-term satellite observational records, especially ~~across a~~on the large scale, ~~still~~ deserves careful investigation. Here, we ~~manifest~~ further confirm that the ~~random forest~~RF, combined with ~~the~~ appropriate ~~covariate~~ covariates exploiting both the spatial and temporal domains, ~~and the~~ together with a model-derived residual calibration module~~, can~~ could be a robust method ~~in~~ for gap-filling of the CCI SM database over China. The superiority of RF-based model in reconstructing SM is further proved by ~~comparing it against the other~~ comparison with four models. ~~Despite this~~Nevertheless, more advanced machine learning strategies, such as deep neural networks (DNN) and long short-term memory (LSTM), are expected to enhance simulation accuracy. ~~Specifically, the ensemble~~ Ensemble approaches that mainly ~~accounting~~ account for the scale biases among different gridded dataset are required. For example, ~~one~~ development of a Bayesian ~~modeling~~ modelling framework that can provide simulation standard error using uncertainty quantification is encouraged (Zhao et al., 2019a).

The variables forcing the proposed model are all ~~from reliable data and~~ available on the long-term~~ worldwide~~ scale globally. Accordingly, ~~the proposed method can~~ our framework could be extended to generate a promising long-term gap-filled SM dataset. This is critical considering ~~the~~ that spatiotemporally continuous SM is demanded for ecological and hydrological research. ~~A promising result with a R^2 of 0.72 is observed when applying our gap-filling approach to long-term SM data sets (2005-2015) in China. In particular, a more accurate trend is achieved relative to that of the original CCI SM when assessed with in situ measurements (0.45 versus 0.23 in terms of R^2). Overall,~~ Thus, the findings of our study ~~may~~ might provide ~~several~~ insights ~~for~~ regarding continuous monitoring of surface water dynamics and drought, and ~~further~~ promote ~~the~~further ~~research~~further research of water resources management and climate change.

Code/Data availability

All the datasets used in this study are open to the public. The National Aeronautics and Space Administration team provides the MODIS products, SRTM DEM data and GLDAS data~~ freely download via the website https://earthdata.nasa.gov/~~. The ESA CCI soil moisture dataset and ERA-5 reanalysis datasets is collected from the European Centre for Medium-range Weather Forecasts (ECMWF)~~ for providing (https://www.ecmwf.int/en/forecasts/datasets)~~. The Brecht Martens, Diego Miralles and their team provides the GLEAM datasets (http://www.gleam.eu/). The China Watershed Allied Telemetry Experimental Research (WATER) project, Chinese Ecosystem Research Network (CERN), and Maqu soil moisture monitoring network provides available in situ measurements at the website http://data.tpdc.ac.cn/en/. The Chinese regional ground meteorological dataset is collected from the National Tibetan Plateau Data Center (http://data.tpdc.ac.cn).

Author contribution

Kai Liu, Xueke Li, and Shudong Wang designed the theoretical formalism. Kai Liu performed the analytic calculations. Both Shudong Wang and Hongyan Zhang contributed to the final version of the paper.

Competing interests

The contact author has declared that neither they nor their co-authors have any competing interests.

Acknowledgments

This study was jointly supported by the Natural Science Foundation of China (42141007 and 41671362), and the Inner Mongolia Autonomous Region Science and Technology Achievement Transformation Special Fund Project (2021CG0045).

References

Almendra-Martín, L., Martínez-Fernández, J., Piles, M., and González-Zamora, Á.: Comparison of gap-filling techniques applied to the CCI soil moisture database in Southern Europe, Remote Sensing of Environment, 258, 112377, https://doi.org/10.1016/j.rse.2021.112377, 2021.

Amani, M., Salehi, B., Mahdavi, S., Masjedi, A., and Dehnavi, S.: Temperature-Vegetation-soil Moisture Dryness Index (TVMDI), Remote Sensing of Environment, 197, 1-14, https://doi.org/10.1016/j.rse.2017.05.026, 2017.

Balsamo, G., Albergel, C., Beljaars, A., Boussetta, S., Brun, E., Cloke, H., Dee, D., Dutra, E., Muñoz-Sabater, J., Pappenberger, F., de Rosnay, P., Stockdale, T., and Vitart, F.: ERA-Interim/Land: a global land surface reanalysis data set, Hydrol. Earth Syst. Sci., 19, 389-407, 10.5194/hess-19-389-2015, 2015.

Belgiu, M. and Drăguţ, L.: Random forest in remote sensing: A review of applications and future directions, ISPRS Journal of Photogrammetry and Remote Sensing, 114, 24-31, https://doi.org/10.1016/j.isprsjprs.2016.01.011, 2016.

Bessenbacher, V., Gudmundsson, L., and Seneviratne, S. I.: Capturing future soil-moisture droughts from irregularly distributed ground observations, Copernicus Meetings, https://doi.org/10.5194/egusphere-egu22-8714, 2022a.

Bessenbacher, V., Seneviratne, S. I., and Gudmundsson, L.: CLIMFILL v0.9: a framework for intelligently gap filling Earth observations, Geosci. Model Dev., 15, 4569-4596, 10.5194/gmd-15-4569-2022, 2022b.

Breiman, L.: Random Forests, Machine Learning, 45, 5-32, 10.1023/A:1010933404324, 2001.

Chen, B., Xu, G., Coops, N. C., Ciais, P., Innes, J. L., Wang, G., Myneni, R. B., Wang, T., Krzyzanowski, J., Li, Q., Cao, L., and Liu, Y.: Changes in vegetation photosynthetic activity trends across the Asia–Pacific region over the last three decades, Remote Sensing of Environment, 144, 28-41, https://doi.org/10.1016/j.rse.2013.12.018, 2014.

Chen, Y., Yang, K., Qin, J., Zhao, L., Tang, W., and Han, M.: Evaluation of AMSR-E retrievals and GLDAS simulations against observations of a soil moisture network on the central Tibetan Plateau, Journal of Geophysical Research: Atmospheres, 118, 4466-4475, https://doi.org/10.1002/jgrd.50301, 2013.

Cristea, N. C., Breckheimer, I., Raleigh, M. S., HilleRisLambers, J., and Lundquist, J. D.: An evaluation of terrain-based downscaling of fractional snow covered area data sets based on LiDAR-derived snow data and orthoimagery, Water Resources Research, 53, 6802-6820, https://doi.org/10.1002/2017WR020799, 2017.

Cui, Y., Yang, X., Chen, X., Fan, W., Zeng, C., Xiong, W., and Hong, Y.: A two-step fusion framework for quality improvement of a remotely sensed soil moisture product: A case study for the ECV product over the Tibetan Plateau, Journal of Hydrology, 587, 124993, https://doi.org/10.1016/j.jhydrol.2020.124993, 2020.

Cui, Y., Zeng, C., Zhou, J., Xie, H., Wan, W., Hu, L., Xiong, W., Chen, X., Fan, W., and Hong, Y.: A spatio-temporal continuous soil moisture dataset over the Tibet Plateau from 2002 to 2015, Scientific Data, 6, 247, 10.1038/s41597-019-0228-x, 2019.

Dente, L., Vekerdy, Z., Wen, J., and Su, Z.: Maqu network for validation of satellite-derived soil moisture products, International Journal of Applied Earth Observation and Geoinformation, 17, 55-65, https://doi.org/10.1016/j.jag.2011.11.004, 2012.

Detto, M., Montaldo, N., Albertson, J. D., Mancini, M., and Katul, G.: Soil moisture and vegetation controls on evapotranspiration in a heterogeneous Mediterranean ecosystem on Sardinia, Italy, Water Resources Research, 42, https://doi.org/10.1029/2005WR004693, 2006.

Dorigo, W., Wagner, W., Albergel, C., Albrecht, F., Balsamo, G., Brocca, L., Chung, D., Ertl, M., Forkel, M., Gruber, A., Haas, E., Hamer, P. D., Hirschi, M., Ikonen, J., de Jeu, R., Kidd, R., Lahoz, W., Liu, Y. Y., Miralles, D., Mistelbauer, T., Nicolai-Shaw, N., Parinussa, R., Pratola, C., Reimer, C., van der Schalie, R., Seneviratne, S. I., Smolander, T., and Lecomte, P.: ESA CCI Soil Moisture for improved Earth system understanding: State-of-the art and future directions, Remote Sensing of Environment, 203, 185-215, https://doi.org/10.1016/j.rse.2017.07.001, 2017.

Dorigo, W. A., Gruber, A., De Jeu, R. A. M., Wagner, W., Stacke, T., Loew, A., Albergel, C., Brocca, L., Chung, D., Parinussa, R. M., and Kidd, R.: Evaluation of the ESA CCI soil moisture product using ground-based observations, Remote Sensing of Environment, 162, 380-395, https://doi.org/10.1016/j.rse.2014.07.023, 2015.

Dorigo, W. A., Wagner, W., Hohensinn, R., Hahn, S., Paulik, C., Xaver, A., Gruber, A., Drusch, M., Mecklenburg, S., van Oevelen, P., Robock, A., and Jackson, T.: The International Soil Moisture Network: a data hosting facility for global in situ soil moisture measurements, Hydrol. Earth Syst. Sci., 15, 1675-1698, 10.5194/hess-15-1675-2011, 2011.

Duan, Z. and Bastiaanssen, W. G. M.: First results from Version 7 TRMM 3B43 precipitation product in combination with a new downscaling–calibration procedure, Remote Sensing of Environment, 131, 1-13, https://doi.org/10.1016/j.rse.2012.12.002, 2013.

Entekhabi, D., Njoku, E. G., Neill, P. E. O., Kellogg, K. H., Crow, W. T., Edelstein, W. N., Entin, J. K., Goodman, S. D., Jackson, T. J., Johnson, J., Kimball, J., Piepmeier, J. R., Koster, R. D., Martin, N., McDonald, K. C., Moghaddam, M., Moran, S., Reichle, R., Shi, J. C., Spencer, M. W., Thurman, S. W., Tsang, L., and Zyl, J. V.: The Soil Moisture Active Passive (SMAP) Mission, Proceedings of the IEEE, 98, 704-716, 10.1109/JPROC.2010.2043918, 2010.

Ford, T. W. and Quiring, S. M.: Comparison and application of multiple methods for temporal interpolation of daily soil moisture, International Journal of Climatology, 34, 2604-2621, https://doi.org/10.1002/joc.3862, 2014.

Fu, G., Crosbie, R. S., Barron, O., Charles, S. P., Dawes, W., Shi, X., Van Niel, T., and Li, C.: Attributing variations of temporal and spatial groundwater recharge: A statistical analysis of climatic and non-climatic factors, Journal of Hydrology, 568, 816-834, https://doi.org/10.1016/j.jhydrol.2018.11.022, 2019.

GCOS: Implementation plan for the global observing system for climate in support of the UNFCCC (2010 update), 2010.

Gruber, A., Scanlon, T., van der Schalie, R., Wagner, W., and Dorigo, W.: Evolution of the ESA CCI Soil Moisture climate data records and their underlying merging methodology, Earth Syst. Sci. Data, 11, 717-739, 10.5194/essd-11-717-2019, 2019.

Gunnarsson, A., Gardarsson, S. M., Pálsson, F., Jóhannesson, T., and Sveinsson, Ó. G. B.: Annual and inter-annual variability and trends of albedo of Icelandic glaciers, The Cryosphere, 15, 547-570, 10.5194/tc-15-547-2021, 2021.

He, J., Yang, K., Tang, W., Lu, H., Qin, J., Chen, Y., and Li, X.: The first high-resolution meteorological forcing dataset for land process studies over China, Scientific Data, 7, 25, 10.1038/s41597-020-0369-y, 2020.

Hu, L., Monaghan, A., Voogt, J. A., and Barlage, M.: A first satellite-based observational assessment of urban thermal anisotropy, Remote Sensing of Environment, 181, 111-121, https://doi.org/10.1016/j.rse.2016.03.043, 2016.

Jing, W., Zhang, P., and Zhao, X.: Reconstructing Monthly ECV Global Soil Moisture with an Improved Spatial Resolution, Water Resources Management, 32, 2523-2537, 10.1007/s11269-018-1944-2, 2018.

Karbalaye Ghorbanpour, A., Hessels, T., Moghim, S., and Afshar, A.: Comparison and assessment of spatial downscaling methods for enhancing the accuracy of satellite-based precipitation over Lake Urmia Basin, Journal of Hydrology, 596, 126055, https://doi.org/10.1016/j.jhydrol.2021.126055, 2021.

Kerr, Y. H., Waldteufel, P., Wigneron, J., Martinuzzi, J., Font, J., and Berger, M.: Soil moisture retrieval from space: the Soil Moisture and Ocean Salinity (SMOS) mission, IEEE Transactions on Geoscience and Remote Sensing, 39, 1729-1735, 10.1109/36.942551, 2001.

Leng, P., Li, Z.-L., Duan, S.-B., Gao, M.-F., and Huo, H.-Y.: A practical approach for deriving all-weather soil moisture content using combined satellite and meteorological data, ISPRS Journal of Photogrammetry and Remote Sensing, 131, 40-51, https://doi.org/10.1016/j.isprsjprs.2017.07.013, 2017.

Li, B., Liang, S., Liu, X., Ma, H., Chen, Y., Liang, T., and He, T.: Estimation of all-sky 1 km land surface temperature over the conterminous United States, Remote Sensing of Environment, 266, 112707, https://doi.org/10.1016/j.rse.2021.112707, 2021a.

Li, L., Dai, Y., Shangguan, W., Wei, N., Wei, Z., and Gupta, S.: Multistep Forecasting of Soil Moisture Using Spatiotemporal Deep Encoder–Decoder Networks, Journal of Hydrometeorology, 23, 337-350, 10.1175/jhm-d-21-0131.1, 2022a.

Li, L., Dai, Y., Shangguan, W., Wei, Z., Wei, N., and Li, Q.: Causality-Structured Deep Learning for Soil Moisture Predictions, Journal of Hydrometeorology, 10.1175/jhm-d-21-0206.1, 2022b.

Li, Q., Li, Z., Shangguan, W., Wang, X., Li, L., and Yu, F.: Improving soil moisture prediction using a novel encoder-decoder model with residual learning, Computers and Electronics in Agriculture, 195, 106816, https://doi.org/10.1016/j.compag.2022.106816, 2022c.

Li, Q., Wang, Z., Shangguan, W., Li, L., Yao, Y., and Yu, F.: Improved daily SMAP satellite soil moisture prediction over China using deep learning model with transfer learning, Journal of Hydrology, 600, 126698, https://doi.org/10.1016/j.jhydrol.2021.126698, 2021b.

Li, X., Zhang, C., Li, W., and Liu, K.: Evaluating the Use of DMSP/OLS Nighttime Light Imagery in Predicting PM2.5 Concentrations in the Northeastern United States, Remote Sensing, 9, 620, 2017.

Li, Y., Piao, S., Li, L. Z. X., Chen, A., Wang, X., Ciais, P., Huang, L., Lian, X., Peng, S., Zeng, Z., Wang, K., and Zhou, L.: Divergent hydrological response to large-scale afforestation and vegetation greening in China, Science Advances, 4, eaar4182, doi:10.1126/sciadv.aar4182, 2018.

Liu, K., Li, X., and Long, X.: Trends in groundwater changes driven by precipitation and anthropogenic activities on the southeast side of the Hu Line, Environmental Research Letters, 16, 094032, 10.1088/1748-9326/ac1ed8, 2021a.

Liu, K., Li, X., and Wang, S.: Characterizing the spatiotemporal response of runoff to impervious surface dynamics across three highly urbanized cities in southern China from 2000 to 2017, International Journal of Applied Earth Observation and Geoinformation, 100, 102331, https://doi.org/10.1016/j.jag.2021.102331, 2021b.

Liu, K., Su, H., Li, X., and Chen, S.: Development of a 250-m Downscaled Land Surface Temperature Data Set and Its Application to Improving Remotely Sensed Evapotranspiration Over Large Landscapes in Northern China, IEEE Transactions on Geoscience and Remote Sensing, 1-12, 10.1109/TGRS.2020.3037168, 2020a.

Liu, K., Wang, S., Li, X., and Wu, T.: Spatially Disaggregating Satellite Land Surface Temperature With a Nonlinear Model Across Agricultural Areas, Journal of Geophysical Research: Biogeosciences, 124, 3232-3251, https://doi.org/10.1029/2019JG005227, 2019.

Liu, Y., Yao, L., Jing, W., Di, L., Yang, J., and Li, Y.: Comparison of two satellite-based soil moisture reconstruction algorithms: A case study in the state of Oklahoma, USA, Journal of Hydrology, 590, 125406, https://doi.org/10.1016/j.jhydrol.2020.125406, 2020b.

Llamas, R. M., Guevara, M., Rorabaugh, D., Taufer, M., and Vargas, R.: Spatial Gap-Filling of ESA CCI Satellite-Derived Soil Moisture Based on Geostatistical Techniques and Multiple Regression, Remote Sensing, 12, 665, 2020.

Long, D., Bai, L., Yan, L., Zhang, C., Yang, W., Lei, H., Quan, J., Meng, X., and Shi, C.: Generation of spatially complete and daily continuous surface soil moisture of high spatial resolution, Remote Sensing of Environment, 233, 111364, https://doi.org/10.1016/j.rse.2019.111364, 2019.

Long, D., Yan, L., Bai, L., Zhang, C., Li, X., Lei, H., Yang, H., Tian, F., Zeng, C., Meng, X., and Shi, C.: Generation of MODIS-like land surface temperatures under all-weather conditions based on a data fusion approach, Remote Sensing of Environment, 246, 111863, https://doi.org/10.1016/j.rse.2020.111863, 2020.

Mao, H., Kathuria, D., Duffield, N., and Mohanty, B. P.: Gap Filling of High-Resolution Soil Moisture for SMAP/Sentinel-1: A Two-Layer Machine Learning-Based Framework, Water Resources Research, 55, 6986-7009, https://doi.org/10.1029/2019WR024902, 2019.

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.

Merlin, O., Jacob, F., Wigneron, J., Walker, J., and Chehbouni, G.: Multidimensional Disaggregation of Land Surface Temperature Using High-Resolution Red, Near-Infrared, Shortwave-Infrared, and Microwave-L Bands, IEEE Transactions on Geoscience and Remote Sensing, 50, 1864-1880, 10.1109/TGRS.2011.2169802, 2012.

Miralles, D. G., Holmes, T. R. H., De Jeu, R. A. M., Gash, J. H., Meesters, A. G. C. A., and Dolman, A. J.: Global land-surface evaporation estimated from satellite-based observations, Hydrol. Earth Syst. Sci., 15, 453-469, 10.5194/hess-15-453-2011, 2011.

Otkin, J. A., Anderson, M. C., Hain, C., Svoboda, M., Johnson, D., Mueller, R., Tadesse, T., Wardlow, B., and Brown, J.: Assessing the evolution of soil moisture and vegetation conditions during the 2012 United States flash drought, Agricultural and Forest Meteorology, 218-219, 230-242, https://doi.org/10.1016/j.agrformet.2015.12.065, 2016.

Prihodko, L., Denning, A. S., Hanan, N. P., Baker, I., and Davis, K.: Sensitivity, uncertainty and time dependence of parameters in a complex land surface model, Agricultural and Forest Meteorology, 148, 268-287, https://doi.org/10.1016/j.agrformet.2007.08.006, 2008.

Ramoelo, A., Cho, M. A., Mathieu, R., Madonsela, S., van de Kerchove, R., Kaszta, Z., and Wolff, E.: Monitoring grass nutrients and biomass as indicators of rangeland quality and quantity using random forest modelling and WorldView-2 data, International Journal of Applied Earth Observation and Geoinformation, 43, 43-54, https://doi.org/10.1016/j.jag.2014.12.010, 2015.

Reichle, R. H., Koster, R. D., De Lannoy, G. J. M., Forman, B. A., Liu, Q., Mahanama, S. P. P., and Touré, A.: Assessment and Enhancement of MERRA Land Surface Hydrology Estimates, Journal of Climate, 24, 6322-6338, 10.1175/jcli-d-10-05033.1, 2011.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat: Deep learning and process understanding for data-driven Earth system science, Nature, 566, 195-204, 10.1038/s41586-019-0912-1, 2019.

Schaake, J. C., Duan, Q., Koren, V., Mitchell, K. E., Houser, P. R., Wood, E. F., Robock, A., Lettenmaier, D. P., Lohmann, D., Cosgrove, B., Sheffield, J., Luo, L., Higgins, R. W., Pinker, R. T., and Tarpley, J. D.: An intercomparison of soil moisture fields in the North American Land Data Assimilation System (NLDAS), Journal of Geophysical Research: Atmospheres, 109, https://doi.org/10.1029/2002JD003309, 2004.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and Freitas, N. d.: Taking the Human Out of the Loop: A Review of Bayesian Optimization, Proceedings of the IEEE, 104, 148-175, 10.1109/JPROC.2015.2494218, 2016.

Sismanidis, P., Bechtel, B., Keramitsoglou, I., Göttsche, F., and Kiranoudis, C. T.: Satellite-derived quantification of the diurnal and annual dynamics of land surface temperature, Remote Sensing of Environment, 265, 112642, https://doi.org/10.1016/j.rse.2021.112642, 2021.

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.

Stroud, J. R., Müller, P., and Sansó, B.: Dynamic models for spatiotemporal data, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63, 673-689, https://doi.org/10.1111/1467-9868.00305, 2001.

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.

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., and Feuston, B. P.: Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling, Journal of Chemical Information and Computer Sciences, 43, 1947-1958, 10.1021/ci034160g, 2003.

Uebbing, B., Forootan, E., Braakmann-Folgmann, A., and Kusche, J.: Inverting surface soil moisture information from satellite altimetry over arid and semi-arid regions, Remote Sensing of Environment, 196, 205-223, https://doi.org/10.1016/j.rse.2017.05.004, 2017.

van Zyl, J. J.: The Shuttle Radar Topography Mission (SRTM): a breakthrough in remote sensing of topography, Acta Astronautica, 48, 559-565, https://doi.org/10.1016/S0094-5765(01)00020-0, 2001.

Wanders, N., Karssenberg, D., de Roo, A., de Jong, S. M., and Bierkens, M. F. P.: The suitability of remotely sensed soil moisture for improving operational flood forecasting, Hydrol. Earth Syst. Sci., 18, 2343-2357, 10.5194/hess-18-2343-2014, 2014.

Wang, A., Lettenmaier, D. P., and Sheffield, J.: Soil Moisture Drought in China, 1950–2006, Journal of Climate, 24, 3257-3271, 10.1175/2011jcli3733.1, 2011.

Wang, C., Xie, Q., Gu, X., Yu, T., Meng, Q., Zhou, X., Han, L., and Zhan, Y.: Soil moisture estimation using Bayesian Maximum Entropy algorithm from FY3-B, MODIS and ASTER GDEM remote-sensing data in a maize region of HeBei province, China, International Journal of Remote Sensing, 41, 7018-7041, 10.1080/01431161.2020.1752953, 2020.

Wang, K., Wang, P., Liu, J., Sparrow, M., Haginoya, S., and Zhou, X.: Variation of surface albedo and soil thermal parameters with soil moisture content at a semi-desert site on the western Tibetan Plateau, Boundary-Layer Meteorology, 116, 117-129, 10.1007/s10546-004-7403-z, 2005.

Wei, F., Wang, S., Fu, B., Brandt, M., Pan, N., Wang, C., and Fensholt, R.: Nonlinear dynamics of fires in Africa over recent decades controlled by precipitation, Global Change Biology, 26, 4495-4505, https://doi.org/10.1111/gcb.15190, 2020.

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, X., Fu, B., Lü, Y., Sun, F., Wang, S., and Liu, M.: Comparison of Four Spatial Interpolation Methods for Estimating Soil Moisture in a Complex Terrain Catchment, PLOS ONE, 8, e54660, 10.1371/journal.pone.0054660, 2013.

Zhang, L., Liu, Y., Ren, L., Teuling, A. J., Zhang, X., Jiang, S., Yang, X., Wei, L., Zhong, F., and Zheng, L.: Reconstruction of ESA CCI satellite-derived soil moisture using an artificial neural network technology, Science of The Total Environment, 782, 146602, https://doi.org/10.1016/j.scitotenv.2021.146602, 2021a.

Zhang, Q., Yuan, Q., Li, J., Wang, Y., Sun, F., and Zhang, L.: Generating seamless global daily AMSR2 soil moisture (SGD-SM) long-term products for the years 2013–2019, Earth Syst. Sci. Data, 13, 1385-1401, 10.5194/essd-13-1385-2021, 2021b.

Zhang, R., Di, B., Luo, Y., Deng, X., Grieneisen, M. L., Wang, Z., Yao, G., and Zhan, Y.: A nonparametric approach to filling gaps in satellite-retrieved aerosol optical depth for estimating ambient PM2.5 levels, Environmental Pollution, 243, 998-1007, https://doi.org/10.1016/j.envpol.2018.09.052, 2018.

Zhang, X., Zhou, J., Liang, S., and Wang, D.: A practical reanalysis data and thermal infrared remote sensing data merging (RTM) method for reconstruction of a 1-km all-weather land surface temperature, Remote Sensing of Environment, 260, 112437, https://doi.org/10.1016/j.rse.2021.112437, 2021c.

Zhang, X., Chen, B., Zhao, H., Fan, H., and Zhu, D.: Soil Moisture Retrieval over a Semiarid Area by Means of PCA Dimensionality Reduction, Canadian Journal of Remote Sensing, 42, 136-144, 10.1080/07038992.2016.1175928, 2016.

Zhao, K., Wulder, M. A., Hu, T., Bright, R., Wu, Q., Qin, H., Li, Y., Toman, E., Mallick, B., Zhang, X., and Brown, M.: Detecting change-point, trend, and seasonality in satellite time series data to track abrupt changes and nonlinear dynamics: A Bayesian ensemble algorithm, Remote Sensing of Environment, 232, 111181, https://doi.org/10.1016/j.rse.2019.04.034, 2019a.

Zhao, W., Duan, S.-B., Li, A., and Yin, G.: A practical method for reducing terrain effect on land surface temperature using random forest regression, Remote Sensing of Environment, 221, 635-649, https://doi.org/10.1016/j.rse.2018.12.008, 2019b.

Zhao, W., Sánchez, N., Lu, H., and Li, A.: A spatial downscaling approach for the SMAP passive surface soil moisture product using random forest regression, Journal of Hydrology, 563, 1009-1024, https://doi.org/10.1016/j.jhydrol.2018.06.081, 2018.

Zhu, X., Liu, D., and Chen, J.: A new geostatistical approach for filling gaps in Landsat ETM+ SLC-off images, Remote Sensing of Environment, 124, 49-60, https://doi.org/10.1016/j.rse.2012.04.019, 2012.

A robust gap-filling approach for ESA CCI soil moisture integrating satellite observations, model-driven knowledge, and spatiotemporal machine learning

Kai Liu[1], Xueke Li[2], Shudong Wang[1,3], Hongyan Zhang[1]

[1]Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China.
[2]Institute at Brown for Environment and Society, Brown University, Providence, RI, 02912, USA
[3]Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters (CIC-FEMD), Nanjing University of Information Science & Technology, Nanjing 210044, China

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

Abstract. Spatiotemporally continuous soil moisture (SM) data are increasingly in demand for ecological and hydrological research. Satellite remote sensing has potential for mapping SM, but the continuity of satellite-derived SM is hampered by data gaps, resulting from inadequate satellite coverage and radio-frequency interference. Therefore, we propose a new gap-filling approach to reconstruct daily SM time series using the European Space Agency Climate Change Initiative. The developed approach integrates satellite observations, model-driven knowledge, and a machine learning algorithm that leverages both spatial and temporal domains. Taking SM in China as an example, the reconstructed SM showed high accuracy when validated against multiple sets of in situ measurements, with root mean square error (RMSE) and mean absolute error (MAE) of 0.09–0.14 and 0.07–0.13 cm^3/cm^3, respectively. Further evaluation with a 10-fold cross validation revealed median values of the coefficient of determination (R^2), RMSE, and MAE of 0.56, 0.025 cm^3/cm^3, and 0.019 cm^3/cm^3, respectively. The reconstructive performance was noticeably reduced both when excluding one explanatory variable and keeping the other variables unchanged, and when removing the spatiotemporal domain strategy or the residual calibration procedure. In comparison with gap-filled SM data based on a satellite-derived diurnal temperature range (DTR), the gap-filled SM data from bias-corrected model-derived DTRs exhibited relatively lower accuracy but higher spatial coverage. Application of our gap-filling approach to long-term SM datasets (2005-2015) produced a promising result (R^2 = 0.72). A more accurate trend was achieved relative to that of the original CCI SM when assessed with in situ measurements (i.e., 0.49 versus 0.28, respectively, in terms of R^2). Our findings indicate the feasibility of integrating satellite observations, model-driven knowledge, and spatiotemporal machine learning to fill gaps in short- and long-term SM time series, thereby providing a potential avenue for applications to similar studies.

1. Introduction

As an essential component of land–atmosphere interactions, soil moisture (SM) substantially impacts the energy, water, and carbon cycles. It plays important roles in hydrological, environmental, and agricultural applications such as evapotranspiration (ET) estimation (Detto et al., 2006), drought assessment (Wang et al., 2011), and flood forecasting (Wanders et al., 2014). SM has been declared by the Global Climate Observing System (GCOS) and United Nations Framework Convention on Climate Change (UNFCCC) as one of the 50 vital variables in terrestrial domains (Gcos, 2010). Availability of spatially and temporally continuous daily all-weather SM data could facilitate improved understanding of ecological and hydrological processes; therefore, provision of a reliable SM dataset is urgently demanded.

Various methods are available for collecting SM data. In situ measurements can capture the temporal variability of SM at the station scale, and many networks designed for such in situ observations have been installed regionally, nationally, and globally, e.g., the crop growth and farmland SM database in China, the North American Soil Moisture Database in North America, and the International Soil Moisture Network (ISMN) (Schaake et al., 2004; Dorigo et al., 2011). Nevertheless, owing to the limited number of ground stations, obtaining spatially continuous SM measurements across large-scale regions remains a challenge. In addition to ground-based observations, SM can be simulated using numerical models. The Global Land Data Assimilation System (GLDAS) and European Centre for Medium-Range Weather Forecasts (ECMWF) fifth-generation global atmospheric reanalysis (ERA5) can model the soil moisture values that have sufficient spatial coverage (Chen et al., 2013; Reichle et al., 2011). However, such model simulations tend to be sensitive to uncertainties related to model structure, forcing, and parameterization (Prihodko et al., 2008; Dorigo et al., 2017).

Satellite observation is considered a powerful technique for retrieving surface SM data, especially given recent improvements in sensor technology. Some SM-dedicated satellites, e.g., the Advanced Microwave Scanning Radiometer-Earth Observation System (AMSR-E), and Advanced Scatterometer (ASCAT) have used the higher C-band and X-band microwave frequencies to collect SM signals. Despite the sensitivity of satellite-derived SM data to atmospheric variability and vegetation coverage, satellites operating with lower L-band radiometers, such as the Soil Moisture and Ocean Salinity (SMOS) (Kerr et al., 2001) and Soil Moisture Active and Passive (SMAP) (Entekhabi et al., 2010), have exhibited great potential for collecting SM data because of the strong capacity of wavelengths in the L-band frequency range to penetrate vegetation. A case worth noting is that the Climate Change Initiative of the European Space Agency (ESA CCI) has generated one set of global SM dataset (Gruber et al., 2019; Dorigo et al., 2017). This CCI SM product blends a series of SM products from active passive microwave satellite sensors, enabling it one complete and consistent observational SM record. Previous studies have revealed reasonable correlation between the CCI SM dataset and in situ measurements obtained over different regions (Dorigo et al., 2015).

The gap issues that remain in current satellite-based SM products relate to a various factors such as radio-frequency interference and orbital changes of the satellite sensors. Considerable effort has been dedicated to filling missing values in satellite-derived SM datasets. Traditional interpolation approaches that are applied to fill gaps rely on the spatial or temporal patterns of the target variable, such as inverse distance weighting and cokriging (Yao et al., 2013; Ford and Quiring, 2014). Other studies (Leng et al., 2017; Llamas et al., 2020; Meng et al., 2021) have focused on the use of statistical methods that mainly depend on the statistical and physical relationships between target variables and explanatory variables. Only recently machine learning strategies have been introduced to the problem of gap-filling in relation to satellite-derived datasets (Zhang et al., 2021b; Zhang et al., 2021a). Such methods have strong capacity for depicting complex relationships of target variables and explanatory variables. For instance, Elsaadani et al. (2021) assessed the spatiotemporal deep learning method for filling the gaps in soil moisture observations, (Li et al., 2022c; Li et al., 2021b) further improved satellite soil moisture prediction using deep learning model. In comparison with statistical-based models, machine learning models might be more flexible and robust, especially with regard to complex scenes and extended coverage (Reichstein et al., 2019).

Most SM gap-filling studies rely on explanatory variables that are required in describing SM dynamics. In addition to satellite-derived vegetation indexes (e.g., normalized difference vegetation index (NDVI) and enhanced vegetation index (EVI)), surface albedo, and land surface temperature (LST), various climatic and geographical factors have been employed in such studies (Almendra-Martín et al., 2021; Cui et al., 2019; Jing et al., 2018). Nevertheless, although appropriate for use in certain regions, most of those variables are less suitable for use in heterogeneous regions and for extended coverage. For example, previous studies (Song et al., 2021; Liu et al., 2020b) that focused on the NDVI and LST tended to achieve better performance in depicting SM in arid and semi-arid regions, but produced unsatisfactory performance in humid areas. Moreover, satellite-derived variables (e.g., optical and thermal infrared parameters) are likely to be impacted by cloud conditions. Accordingly, researchers have attempted to explore effective information for promoting model establishment and application. Some studies used the feature transform approach to extract distinct signals for driving models. Principal component analysis (PCA) and wavelet decomposition have been employed to reconstruct SM and other satellite-based parameters (Uebbing et al., 2017; Almendra-Martín et al., 2021). Despite reasonable model performance achieved in humid and semi-arid regions (Zhang et al., 2016; Almendra-Martín et al., 2021), some studies found no substantial improvement in model performance in areas of cropland in semi-humid regions when using the PCA (Wang et al., 2020). Other studies focused on the distinct dataset source for gap-filling models. Soil moisture from GLDAS, ERA5, China Meteorological Administration Land Data Assimilation System (CLDAS) and Fengyun Microwave Radiation Imager is considered (Long et al., 2019; Cui et al., 2020). The gap-filling models integrating these unique dataset sources are able to describe SM dynamics, but uncertainties remain in relation to humid regions and areas subject to the freezing-thaw process (Song et al., 2021; Cui et al., 2019). Overall, progress regarding the availability of explanatory variables for use in models for reconstruction of SM is inadequately and this is especially critical for machine learning gap-filling models that are sensitive to the structure of the input sequences (Mao et al., 2019).

Although earlier studies focused on completing SM datasets, most partially addressed a specific case of satellite observations but failed to consider larger continental regions. Almendra-Martín et al. (2021) and Liu et al. (2020b) applied reconstruction algorithms to the CCI SM product in regional Europe and Oklahoma, USA, respectively, and Cui et al. (2019) continuously promote this approach in the Tibetan Plateau. Such models rely on machine learning algorithms and a variety of satellite-based variables. Furthermore, research on the challenging case of SM time series at the daily scale(Zhang et al., 2021b; Long et al., 2019), which is fundamental to the exploration of SM dynamics, and the quantification of the associated impact on the contribution to climate change and the water cycle is limited (Bessenbacher et al., 2022a).

Here, we propose a robust gap-filling methodology for reconstruction of a spatially continuous daily ESA CCI SM dataset, primarily based on satellite observations, model-driven knowledge, and one spatiotemporal random forest algorithm. Our model was tested by application to continental China, which has suitable variability in terms of landscape and climatic conditions. Specifically, the feasibility and merit of the developed model were demonstrated by the following: 1) evaluation of the gap-filled results using in situ measurements and holdout cross validation, and comparison against those of other models, and 2) examination of model uncertainty in terms of the filtered explanatory variables, and consideration of the extension of the proposed model to one long-term period.

2. Study region

China is located from 3°51′N to 53°33′N and from 73°33′E to 135°05′E, covering an area of approximately 9.6×10^6 km^6 (Fig. 1). A variety of terrain types are presented across China, including the plain, basin, plateau, mountain and hill. These diverse terrains inevitably result in noticeable spatial differences in precipitation and temperature, accompanying the elevation decreasing from west to east. Seven climate zones can be identified in China, including arid, semi-arid, arid/semi-wet, wet/semi-arid, wet, moist, and over-wet climates. The identification of this zoning system is based on a China's humidity index map produced by the National Earth System Science Data Center, National Science & Technology Infrastructure of China (http://www.geodata.cn).

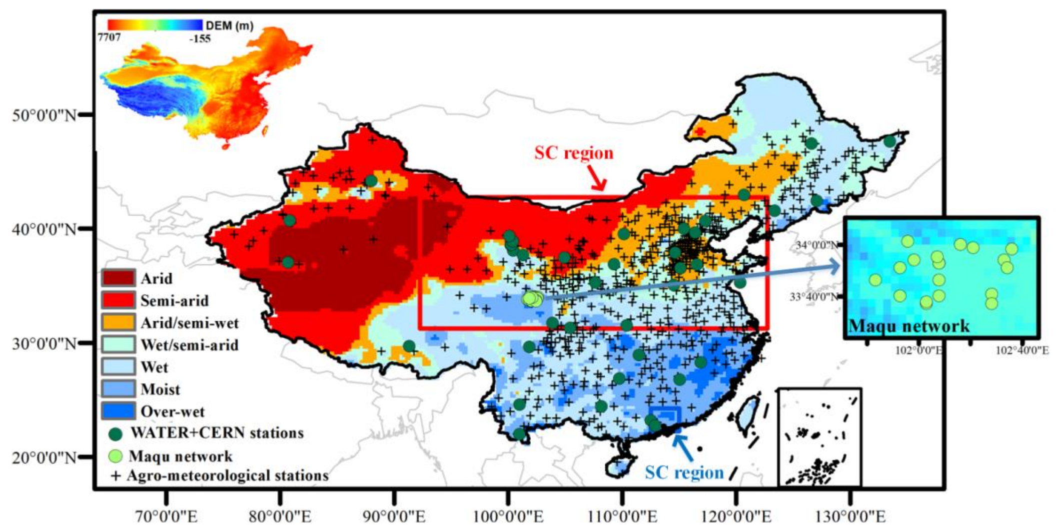


Figure 1: The study region and the selected in situ soil moisture sites. The figure in the upper-left corner shows the Digital Elevation Model (DEM) information. The detailed distribution of dense in situ measurements in the Maqu network is shown in the figure on the far right. Two regional areas for uncertainty analysis (i.e., northern China (NC) and southern China (SC)) are bordered by the rectangles.

3. Materials and methods

The object of this study was to reconstruct CCI SM data gaps to produce spatially continuous data records. The basic principle of the proposed gap-filling approach is to efficiently determine the correlation between SM records and the corresponding explanatory variables, which can be expressed as follows:

$$SM = f(V_1, V_2, V_3 \cdots\cdots V_k) + \varepsilon, \tag{1}$$

$$V_i \in R^{N,T}, \tag{2}$$

where SM is the soil moisture, V_i is the corresponding explanatory vectors, and k is the number of the input variables. V_i can be a vector, and the sample number is determined the spatial domain (N) and temporal domain (T). f is one function that can be either linear or nonlinear. ε represents the model residual. In machine learning ensemble, f represents a black box model that does not have one specific form.

The proposed methodology involves three core steps: (i) using a regression subset selection approach and a variable correction procedure to filter explanatory variables from the satellite observations and model-driven knowledge, and to correct the systematic variable bias between them; (ii) training a machine learning algorithm to determine the SM-explanatory variables correlation based on the selected optimal parameters and the available pixels identified with a spatiotemporal window search strategy, and then applying the established correlation to retrieve the unavailable SM pixels; and (iii) conducting geographically weighted regression and Gaussian filtering to calibrate the model-derived residuals.

Figure 2 shows a schematic of the overall procedure including the dataset processing, model implementation and model analysis.

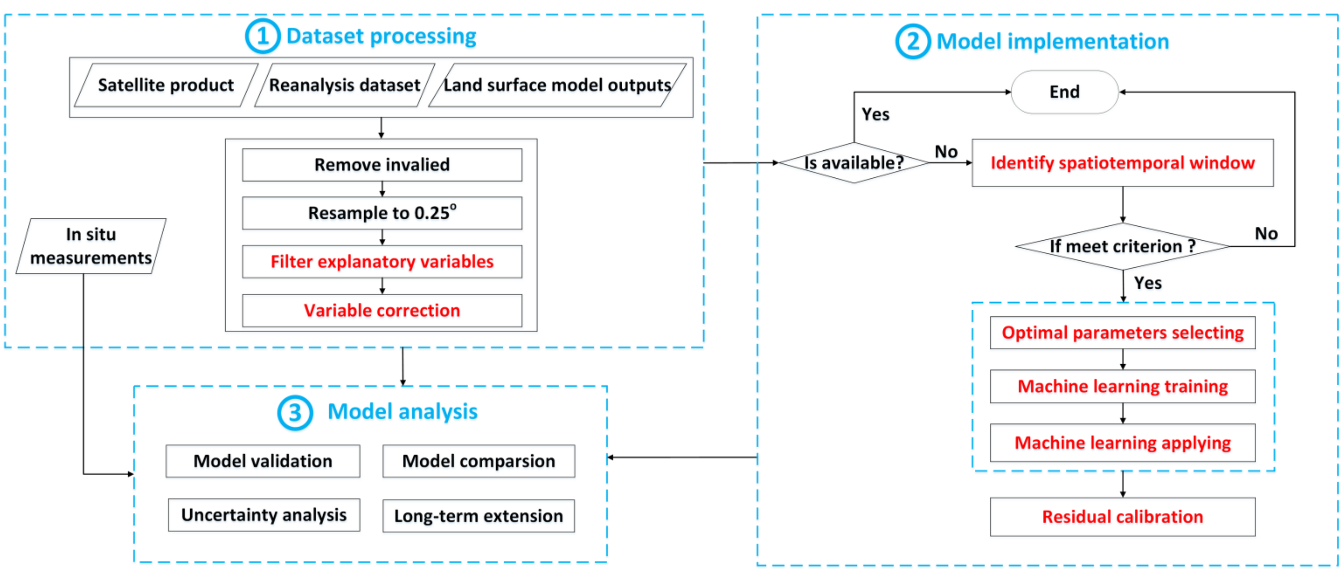


Figure 2: The schematic of overall procedure. The red text denotes the core procedures conducted in the proposed model, which will be described in the following sections.

3.1 Dataset processing

The dataset used includes satellite product, reanalysis dataset, land surface model outputs and in situ measurements (Table 1 and Table S1). Details about these datasets are described in the following sections.

Table 1. Summary of the dataset used for the proposed model. Other dataset for the preliminary analysis but not the final utilization of the model is exhibited in supplementary Table S1.

ID	Variables	Source	Resolution (spatial/temporal)
1	Soil moisture	ESA CCI	0.25°/daily
2	Surface albedo	MCD43C3	0.05°/16 day
3	NDVI	MOD13C1, MYD13C1	0.05°/16 day
4	Land surface temperature (LST)	MYD11C1	1km/instantaneous
5	Precipitation	China Meteorological Forcing Dataset	0.1°/3 hourly
6	Potential evapotranspiration (PET)	GLEAM	0.25°/daily
7	Soil moisture	ERA5	0.25°/hourly
8	Land cover classification	MCD12Q1	500/annual
9	Digital Elevation Model (DEM)	SRTM	90m
10	Surface temperature	Noah simulations from previous work	1km/3 hourly
11	Surface temperature	ERA5	0.25°/hourly
12	Surface temperature	GLDAS	0.25°/3-hourly
13	Soil moisture	GLDAS	0.25°/3-hourly
14	Soil moisture	GLEAM	0.25°/daily
15	in situ soil moisture	China Watershed Allied Telemetry Experimental Research (WATER)	daily
16	in situ soil moisture	Chinese Ecosystem Research Network (CERN)	5 day
17	in situ soil moisture	Tibetan Plateau observatory of plateau scale soil moisture and soil temperature (Tibet-Obs)	daily
18	in situ soil moisture	China's agrometeorological observation network	10 daily

3.1.1 Satellite product

The ESA CCI SM dataset is provided by the Climate Change Initiative program of the European Space Agency. This product is primarily composed of three types of daily dataset sources, i.e., active, passive, and active-passive combined microwave products (Dorigo et al., 2017). Despite the wide spatiotemporal coverage of CCI SM, the data gap remains a major challenge that hampers its further application. Here, we select the daily combined microwave products version 4.5, with a spatial resolution of 0.25°. The inconsistent data in the CCI combined SM are filtered using the quality flag variable.

A variety of Moderate Resolution Imaging Spectroradiometer (MODIS) products are collected, including the daily LST (MYD11C1), 16-day composite albedo (MCD43C3) and vegetation indices, i.e., normalized difference vegetation index (NDVI) and enhanced vegetation index (EVI), and the 8-day composite leaf area index (LAI) (MCD15A2H). All these datasets are collected at MODIS 6 collection. We calculate the diurnal temperature range (DTR) by subtracting the night LST from daytime LST. The NDVI and EVI are averagely obtained from the two products: MOD13C1 and MYD13C1. All the selected products are screened out using the quality variables to maintain only the available pixels with good quality. We also collect the 0.05° annual land cover product (MCD12Q1) for quality control of CCI SM.

We use the Digital Elevation Model (DEM) dataset provided by NASA's Shuttle Radar Topography Mission (SRTM) (Van Zyl, 2001) to retrieve several relevant topographic metrics, including slope, aspect, and the topographic position index (TPI) (Guisan et al., 1999). The TPI is calculated by subtracting the focal grid elevation from the mean elevation of the eight surrounding grids. The TPI is potentially correlated better with surface variables such as snow depth and SM in comparison with the DEM (Cristea et al., 2017). Positive (negative) TPI values mean that the target grid is higher (lower) than the average of its surroundings.

Considering the low accuracy of satellite SM for snow-covered pixels, pixels that have both daytime LST lower than 0 °C and albedo higher than 0.3 are removed (Cui et al., 2020). We also remove pixels for which a water body accounts of more than 20% of the total area. To overcome the spatial resolution differences among the diverse products available, all datasets are resampled to 0.25° spatial resolution by averaging the pixel values.

3.1.2 Reanalysis dataset and land surface model outputs

We collect the soil moisture data from ERA5, a global atmospheric reanalysis dataset released by the ECMWF (Balsamo et al., 2015). The data assimilation system used for ERA5 is the ECMWF Integrated Forecast System (IFS), and the meteorological forcing for retrieving soil moisture is from the ERA atmospheric reanalysis. Here we select the daily averaged SM from the first soil layer (0–7 cm) to match with satellite CCI SM.

Daily potential evapotranspiration (PET) and surface soil moisture (0–15 cm) is collected from the Global Land-surface Evaporation Amsterdam Methodology (GLEAM) dataset. GLEAM is based on a general land surface model that focuses on soil moisture and evapotranspiration (Miralles et al., 2011). PET in the GLEAM is calculated with the Priestley–Taylor formula based on multiple reanalysis datasets, while the soil moisture is calculated with a soil-water module based on water cycle balance.

Four meteorological variables, i.e., precipitation, air temperature, solar radiation and wind, are obtained from the China Meteorological Forcing Dataset. This dataset is generated through fusion of in situ station data, remote sensing products, and reanalysis datasets (He et al., 2020). Considering the lag effect of precipitation on surface water dynamics, we use the five-day antecedent precipitation (AP) to replace the daily precipitation (Wei et al., 2020).

Three surface temperature sources are additionally collected for uncertainty analysis. Two sources are collected from the ERA5 and GLDAS ensemble model. Considering the model uncertainties caused by regional surface characteristics and climatic conditions, we simulate surface temperature and surface SM (0–10 cm) by implementing a Noah model that is forced with meteorological variables from the Chinese regional ground meteorological dataset and the surface condition parameters from MODSI. This dataset is previously used in our work (Liu et al., 2020a; Liu et al., 2021b).

3.1.3 In situ measurements

A variety of spatially sparse in situ soil moisture measurements is collected to evaluate the accuracy of gap-filled SM. We collect in situ soil moisture observations at 39 sites obtained from the China Watershed Allied Telemetry Experimental Research (WATER) project and the Chinese Ecosystem Research Network (CERN). These validation stations are set up in a relatively large homogeneous area dominated by vegetation covers (cropland, woodland and grassland) or desert lands. In addition, 657 in situ soil moisture measurements covered by cropland are collected from the Chinese agro-meteorological and ecological observation network.

We also collect the dense in situ measurements at the Maqu soil moisture monitoring network. The Maqu network (33°30′–34°15′N, 101°38′–102°45′E) is located on the north-eastern border of the Tibetan Plateau (Fig. 1) (Dente et al., 2012). In this network, 20 sites are distributed over a uniform grassland cover, located in the large valley of the Yellow River. Maqu network has demonstrated strong capability in monitoring the spatial and temporal SM variability with high accuracy (Su et al., 2013; Wei et al., 2019). The locations and detailed information of all available sites are displayed in Fig.1 and Table S2.

3.1.4 Filter explanatory variables

Explanatory variables related to atmospheric, geophysical, ecological, and hydrological variables are conducive to capturing SM variability. The significance percentage produced by the regression subset selection model (Fu et al., 2019; Liu et al., 2021a) is employed to measure the impacting probability of the explanatory variables, where a high significance percentage indicates strong capability in depicting SM (details in Text S1). We conducted the subset selection model analysis based on a dataset from 2005 to 2015, and 15 variables were selected as input parameters, including seven surface environmental variables, i.e., albedo, NDVI, EVI, LAI, DTR, PET and ERA SM, three elevation variables, i.e., TPI, aspect and slope, three climatic variables, i.e., AP, air temperature, wind, and two geographical factors, i.e., latitude and longitude. All the variables are available from reliable datasets at the continental scale. Gaps presented in these variables were not considered further to avoid introducing additional errors.

As illustrated in Fig. 3(a), albedo, NDVI, EVI, LAI, DTR, AP, PET, ERA SM, TPI, and air temperature have the highest significance percentage in terms of correlation with CCI SM. We excluded aspect, slope, wind, latitude, and longitude owing to their low correlations with SM. The EVI, NDVI, and air temperature were also not considered in further application because the EVI and LAI are closely correlated with NDVI, and air temperature is strongly correlated with DTR. All the selected covariates are physically meaningful in depicting SM. Specifically, the atmospheric variables (i.e., precipitation and PET) are suitable for capturing the temporal dynamics of SM, and the topographic variables are included both to depict the orographic effects and to recapture the spatial pattern of SM. DTR exhibits correlation with SM owing to its capacity in taking account of land-atmosphere coupling. ERA surface moisture was also included to reproduce satellite SM.

To verify the results based on the regression subset selection model, we employed the permutation feature importance to measure the relative importance of each predictor variable. Consistent patterns between the significance percentage and permutation importance further indicate the feasibility of the selected variables in modelling SM. Additionally, because these variables are derived from optical remote sensing, reanalysis datasets, and land surface model products, they have potential for extension to large regions owing to their high availability (Fig. 3(b)).

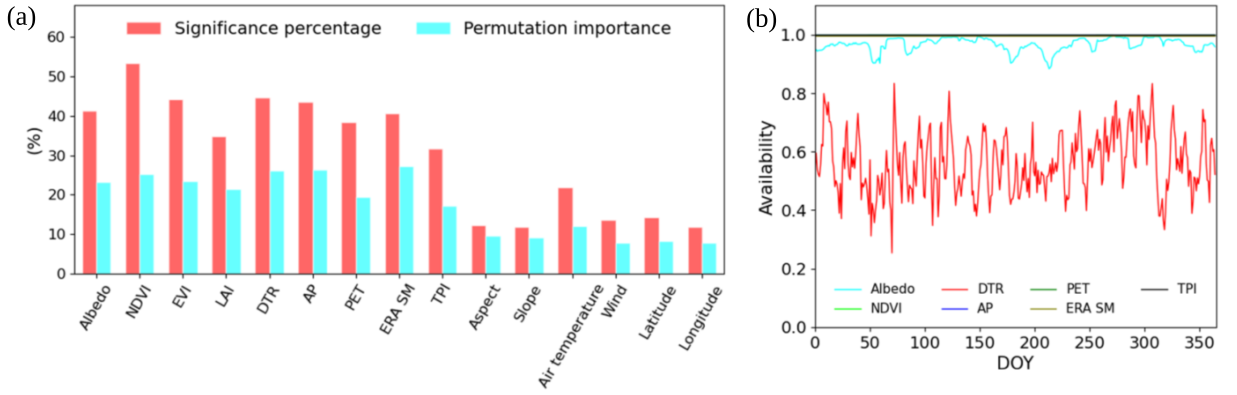


Figure 3: The correlation and availability of dataset used. (a) The significance percentage and permutation importance of the selected variables in correlation to CCI SM. (b) The availability of the selected variables.

3.1.5 Variable correction

Systematic biases are unavoidable in reanalysis datasets and land surface model outputs, and these biases can be propagated in dynamic modeling. Accordingly, bias correction is required prior to the gap-filling procedure to ensure a consistent simulated output. Specifically, to make the modeled values (i.e., ERA SM) comparable with the satellite observations (i.e., ESA CCI SM), we used a correction procedure that primarily combines a variance scaling algorithm and a linear scaling algorithm (Long et al., 2020; Zhang et al., 2021c). The used procedure can be illustrated with the following equations:

$$\begin{cases} SM_{c1} = SM_{ERA}(t_{av}) - \mu\big(SM_{ERA}(t_{av})\big) + \mu\big(SM_{ESA}(t_{av})\big) \\ SM_c = \mu(SM_{c1}) + (SM_{c1} - \mu(SM_{c1})) \times \sigma\big(SM_{ESA}(t_{av})\big)/\sigma\big(SM_{c1} - \mu(SM_{c1})\big) \end{cases} \tag{3}$$

where SM_{ERA} is the raw ERA SM time series of the target grid pixel; t_{av} is time series in which pixels the object grid are available; SM_{ESA} is the ESA SM of the grid; μ and σ are the mean value and the standard deviation, respectively. SM_c is the corrected ERA SM that is assumed to have a spatial pattern (i.e., consistent means and standard deviations) with the CCI SM. In our study, a dataset comprising time series from 2005 to 2015 was used to conduct the correction procedure to guarantee sufficient samples. Examples illustrating the performance of the ERA SM correction can be found in Fig. S1. Despite being conducted on SM, this calibration procedure could be applied to other parameters (e.g., DTR) when replaced with numerical model outputs.

3.2 Model implementation

3.2.1 Machine learning regression

Despite being easy to implement and requiring less computational resources, traditional regression-based methods such as generalized linear models and multivariate regression splines generally insufficiently consider the probability density functions in assessing model performance. Machine learning approaches could be much more flexible than conventional parametric models owing to their ability to handle nonlinear relationships and complex interactions. Among the machine various learning models, the random forest (RF) algorithm, acting as an enhanced decision tree model, is an effective and powerful tool in interpreting earth variables(Belgiu and Drăguţ, 2016). As illustrated in Fig. 4(a), RF is a hierarchical tree diagram that is based on a nonparametric strategy and has the capacity to add a variety of parameter layers into the model (Breiman, 2001). This decision tree model is composed of many nodes and edges within each tree structure, mainly including two types of nodes: split nodes and leaf nodes. The split node is related to a test function that is employed to split the input data, whereas the leaf node is associated with the final decision. Unlike the standard decision tree model that relied on the whole data set, RF trains each tree on bootstrap resamples. This model only considers the randomly selected variables rather than the total variables. By this means, the outcome is decided by a majority voting or averaging strategy.

In this study, the RF model is implemented using the function 'RF Regressor' from the Python Library (Shahriari et al., 2016). Specifically, the built-in functions are used to assess the importance of each covariate by using the out-of-bag samples. We use the 'Bayesian Optimization' module (http://rmcantin.github.io/bayesopt/html/bopttheory.html) to select the best hyperparameters in driving RF algorithm. Four critical parameters deciding the RF algorithm include the number of trees (n_estimators), the maximum tree depth (max_depth), the minimum number of samples for splitting an internal node (min_samples_split), and the number of features (max_features). For each specific climate region, the Bayesian optimization process is carried out within 20 iterations to optimal parameters. The training procedure is mainly based on the dataset covering 2003—2008. Optimal parameters in the seven climate regions are listed in Table S3.

3.2.2 Identify spatiotemporal window

One critical issue relate to the machine learning model is how best to efficiently explore the informative covariates. Here, we use a spatiotemporal strategy to capture the spatial and temporal SM and the related covariate dynamics. Our strategy primarily relies on the available pixels within a regional subset, thereby allowing more pixels of interest to participate in the regression. Figure 4(b) provides the diagram of the spatiotemporal window search strategy.

An adaptive strategy is employed to determine the optimal spatiotemporal window size. Two critical variables are adopted to identify the window size, i.e., the size of the spatial window (sw) and the number of temporal days (nd). To find the optimal sw and nd, we continually increase the value of sw and nd from the initial values until the samples participating for regression meet the criterion, i.e., the number of available pixels within the searched window should be no less than eight times of the participating explanatory variables (i.e., seven) (Svetnik et al., 2003; Liu et al., 2020a). Here an initial sw is set to 5 and an initial nd is set to 1. Considering that a fraction of gaps occur in the satellite dataset (e.g., LST and albedo) and the optimal window may not exist, the maximum values of sw and nd are introduced to terminate this process. A sensitivity analysis is conducted with the independent dataset to select the two maximum values. Specifically, we conduct a cross validation during 2003—2008 to evaluate the accuracy of the gap-filling model. The increasing maximum nd from 1 to 7 with intervals of length 1 is tested, and the maximum sw is tested from 4 to 10 with intervals of length 1. The values that yield the lowest RMSE (Fig. 4(c)) are selected, and finally, we set maximum sw to 7 and the maximum nd to 4. Note that we also conduct sensitivity analysis for each climate region and find no substantial differences in the resulting optimal values of two parameters among seven climate regions. This is probably because this sensitivity analysis is more reliant on model structure rather than sample characteristics.

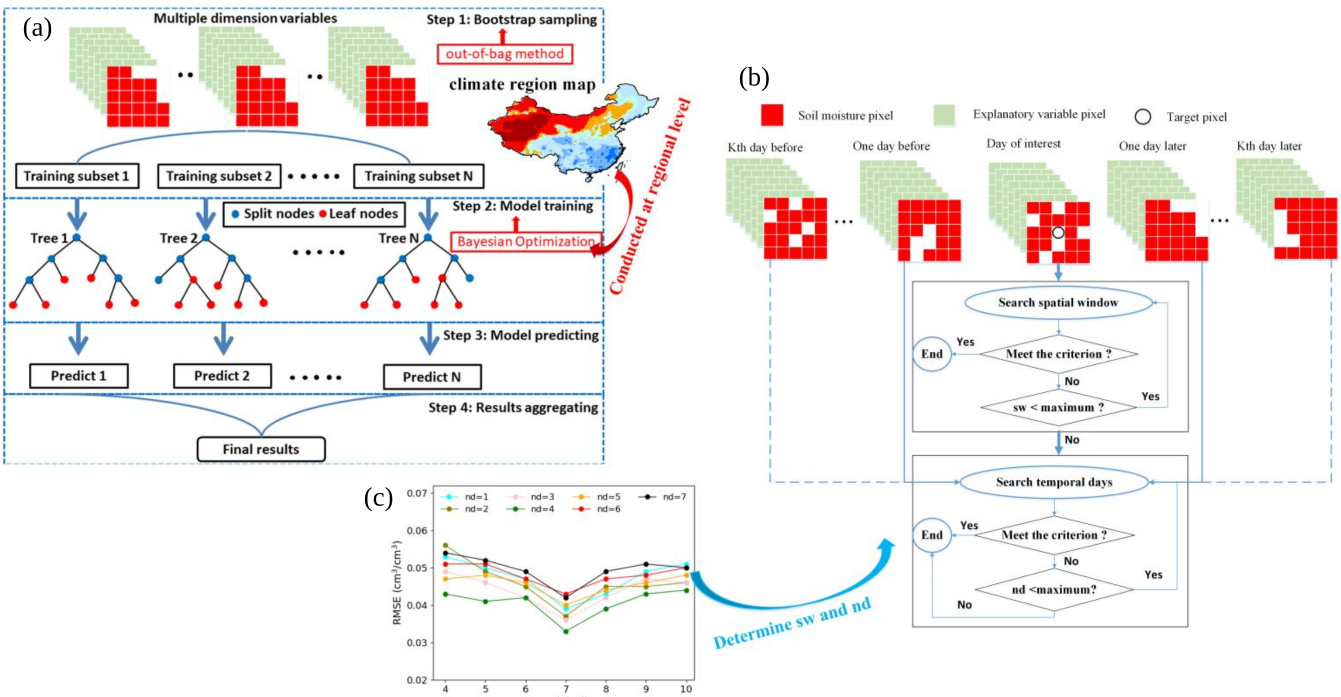


Figure 4: (a) The diagram of the random forest model implemented for a multidimensional dataset. (b) The diagram of spatiotemporal window determination strategy for random forest regression. (c) The results of the sensitive analysis regarding two maximum values, i.e., the size of the spatial window (sw) and the number of temporal days (nd), for terminating the searching process.

3.2.3 Residual calibration

Considering that the machine learning model might not fully account for the variability in SM, the original reconstruction needs to be calibrated, which can potentially remove the bias resulting from neglected variables such as those are excluded for model establishment (Zhu et al., 2012; Liu et al., 2020a). In practice, we add the interpolated model residuals to the original reconstructions. The geographically weighted regression (GWR) model, which is an extension of the traditional linear regression model (Li et al., 2017), is applied to interpolate the RF-derived residuals. This procedure is based on the samples within the searched window for each target pixel. The model residual (ε_j) derived from Eq. (2) can be described using the explanatory variables as follows:

$$\varepsilon_j = \beta_0\left(u_j, v_j\right) + \sum_{i=0}^{k} \beta_i\left(u_j, v_j\right) X_{ij}, \tag{4}$$

where $\beta_0\left(u_j, v_j\right)$ and $\beta_i\left(u_j, v_j\right)$ are the regression coefficients estimated at the jth pixel, and $\left(u_j, v_j\right)$ are the coordinates. The regression coefficients can be estimated using the observations within the self-adaptive searched window as follows:

$$\begin{cases} \hat{\beta}\left(u_j, v_j\right) = \left(X^T\left(W\left(u_j, v_j\right)\right)X\right)^{-1} X^T W\left(u_j, v_j\right) Y \\ \qquad\qquad w_{ij} = \left[1 - \left(d_{ij}/b\right)^2\right]^2 \end{cases}, \tag{5}$$

where $\hat{\beta}\left(u_j, v_j\right)$ is the coefficient matrix composed of coefficients from each explanatory variable; X and Y are the explanatory variable matrix and the dependent variable (i.e., SM) vector, respectively. Here latitude, longitude and seven explanatory variables selected are used to implement the GWR model. $W\left(u_j, v_j\right)$ is the weight matrix composed of w_{ij}, d_{ij} is the Euclidean distance between the observation ith and the jth point, a and b is the window radius.

Before adding to the original reconstruction, the GWR interpolated residual is further smoothed with a normalized k × k Gaussian filter with a standard deviation of σ. This procedure can remove the grid-like artifacts that extensively exist in statistical model outcomes. Base on the optimization procedure (Sismanidis et al., 2021; Liu et al., 2019), we set k = 5 and σ =1.5.

3.3 Model analysis

3.3.1 Model validation

Model validation was conducted using data from 2009 when sufficient number of ground measurements were collected. The top layer SM measurements from the in situ stations were first used to evaluate the accuracy of the reconstructed results. Considering the scale mismatch between the sparse distribution of in situ stations and the CCI SM product (~25 km), we used the Disaggregation based on Physical And Theoretical scale Change (DISPATCH) model (Merlin et al., 2012) to disaggregate the 0.25° reconstructions to 1-km resolution. Detailed descriptions regarding this disaggregation method can be found in Supplementary Text S2.

Evaluating the gap-filled SM with in situ measurements is supposed to produce biases that can be caused by scale mismatching and disaggregation model performance. To account for this, holdout cross validation with 10 replicates was performed in 2009 to evaluate the model accuracy. For each replicate, we randomly held out 10% of the pixels, that is manually introducing gaps for these pixels, and trained the model with the remaining 90% of the dataset. Specifically, the pixels during all periods were first rearranged into a time series and then 10% of them were dropped in each replicate. After the gap-filled SM series of hold-out pixels were reconstructed from the training set, they were validated against the original SM.

To reveal the physical plausibility of gap-filled SM, we paid particular attention to the evaluation of gap-filling SM under extremely dry conditions. Extreme drought is defined based on meteorological condition, that is, the Palmer Drought Severity Index (PDSI) of less than −2 over 8 consecutive months or longer (Fig. S2).

The statistics used for the model accuracy assessment include the coefficient of determination (R^2), the root mean square error (RMSE), the mean absolute error (MAE), the average error bias (BIAS), and the unbiased RMSE (ubRMSE). In addition, Nash-Sutcliffe Efficiency (NSE) is used to measure the overall performance of the proposed model. All these metrics have been extensively used for evaluating satellite SM.

3.3.2 Model comparison

The proposed method was compared against four extensively used models that adopt the same explanatory variables and spatiotemporal window search strategy. The first one is the conventional multiple linear regression (MLR) approach. Three typical machine learning approaches, i.e., Extreme gradient boost (XGB), Support vector machine (SVM) and Artificial Neural Network (ANN), are also used for comparison. Detailed descriptions of four available models can be found in supplementary Text S3.

3.3.3 Uncertainty analysis

Considering the criticality of explanatory variables in simulating SM, uncertainty analyses regarding these selected variables were conducted. We first investigated the accuracy of the reconstruction model that excludes one participating variable. Given the critical importance of satellite-derived DTR and the severe issues of missing data in satellite-observed LST products, we further investigated the substitution performance of other surface temperature sources in reconstructing SM, i.e., i.e., Noah, ERA and GLDAS. This analysis was conducted by focusing on two regions (in Fig. 1) that have sufficient data sources to support our experiments (Liu et al., 2020a; Liu et al., 2021b): one region is in northern China covering mostly arid and semi-arid areas, while the other region is in southern China covering mostly wet areas.

Since the reanalysis SM is a vital input in our approach, we also compare it with the other two products to evaluate the feasibility of ERA data in reconstructing CCI SM. GLEAM and Noah surface SM are respectively employed to replace the ERA SM while other explanatory variables keep the rest the same.

3.3.4 Long-term extension

The available dataset forcing for our model has a long record, indicating potential for modelling long-term SM products. To verify this, the proposed gap-filling method was further extended to the long-term ECA CCI SM databases of 2005—2015. We also investigated the trend of the SM series during this period, which was obtained via Sen's slope and M-K significance analysis (Li et al., 2021c; Liu et al., 2021a). The trends from the reconstructed SM series were also compared with those from the original CCI SM, which were evaluated against in situ measurements.

4. Results and discussion

4.1 Spatiotemporal patterns

The spatiotemporal pattern of the original daily CCI SM and the corresponding gap-filled dataset in 2009 is first checked. As shown in Fig. 5(a) (and (Fig. S3)), a considerably large gap occurs in the original CCI SM, and this gap problem is greater in winter. We reconstruct the contaminated SM pixels using the spatiotemporal RF model. Most of the contaminated pixels (more than 85%) are reconstructed. Relatively few missing pixels are gap-filled in winter in comparison with other seasons, primarily because of the heavy contamination of clear pixels caused by frequent occurrence of cloud during this period. It means that the learning capacity of the spatiotemporal machine learning method is constrained when encountering limited satellite observations.

Figure 5(b) shows the boxplot of the original versus gap-filled SM on selected days in 2009. Conformity exists between the original and reconstructed SM for most days. A similar pattern in variance and magnitude is also observed for the SM of the monthly average and the selected days, as illustrated in Fig. 5(c); that is, large difference occurs in winter and spring. This can be attributed to the fact that the original CCI SM provides fewer training data from October to May of the following year. Additionally, the distribution of CCI SM is more uneven in this period, which might reduce model performance owing to the limited representation of training samples (Stroud et al., 2001).

In terms of different climate regions, minor discrepancy is evident between the original and the reconstructed SM (Figure 5(d)), with bias in the median SM values of less than 8%. It means that the reconstructed SM has strong variation depicting capacity. Small overestimation occurs in arid regions, which originally have less soil water storage.

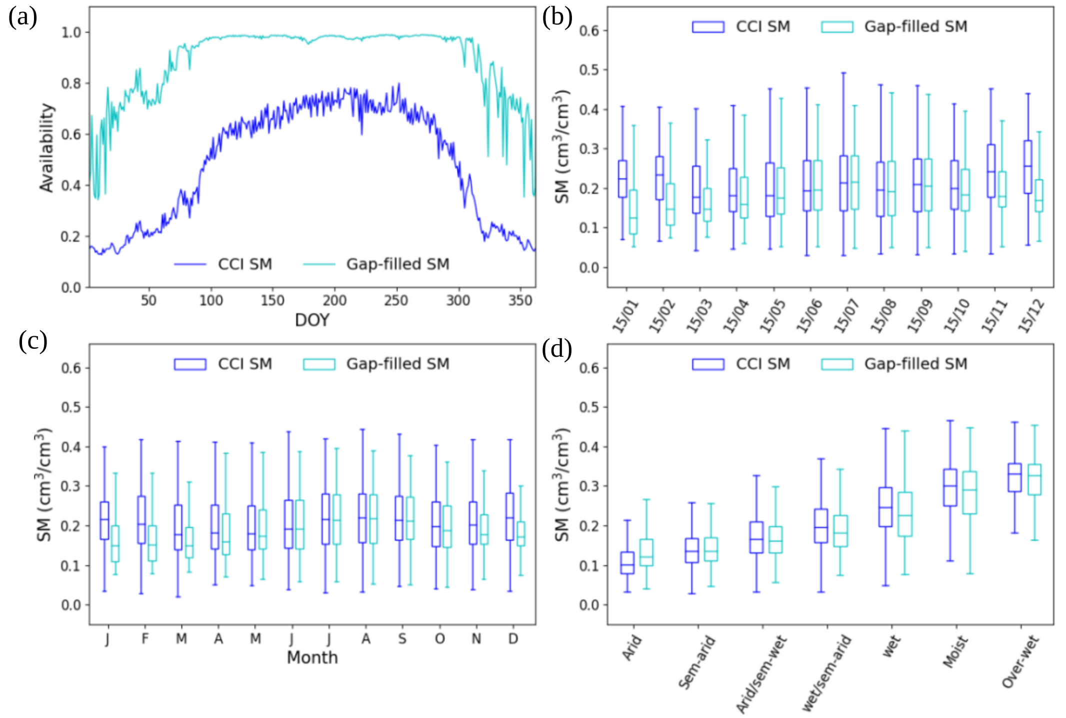


Figure 5: The comparison between CCI dataset and gap-filled SM in 2009. (a) The plots of the availability of CCI dataset and gap-filled SM. (b) The boxplot of the CCI dataset and gap-filled SM on the selected days. (c) The boxplot of month-average CCI and gap-filled SM. (d) The boxplot of raw and gap-filled SM regarding seven climate regions.

Figure 6 exhibits the spatial distributions of the original CCI SM and the reconstructed SM on selected days in 2009. The humid regions are mostly concentrated in southern China adjacent to the coast of the western Pacific, whereas the dry regions are mainly distributed in northern and western parts of China. A considerable fraction of contaminated pixels is observed on the selected days, and this contamination is severe in winter season and in mountainous areas (e.g., Tibet Plateau and Mongolian Plateau). Almost all the contaminated pixels from March to October are reconstructed; meanwhile, the proposed model reconstructs the most contaminated pixels for the remaining months. Owing to the additional valid values provided by gap-filled pixels, more spatial variation is depicted in the reconstructed SM images. Missing pixels still occur in the reconstructed SM images especially in the cold seasons. This can be related to the fact that the surface temperature, ET, and precipitation are more connected in the warm season through energy balance considerations and atmospheric circulation. Some of these invalid pixels correspond to snow- and water-covered regions that have been removed beforehand. Because missing earth data are to a large extent not at random, statistical measures of comparative analysis among them tends to produce bias (Bessenbacher et al., 2022b). To account for this, paired histograms of two datasets are compared to explore the value distribution properties. The histograms show the gap-filled dataset does not impact the SM distribution in warm seasons, that is, in agreement with the CCI dataset. There is also noticeable bias in cold seasons, especially in the very low range of SM.

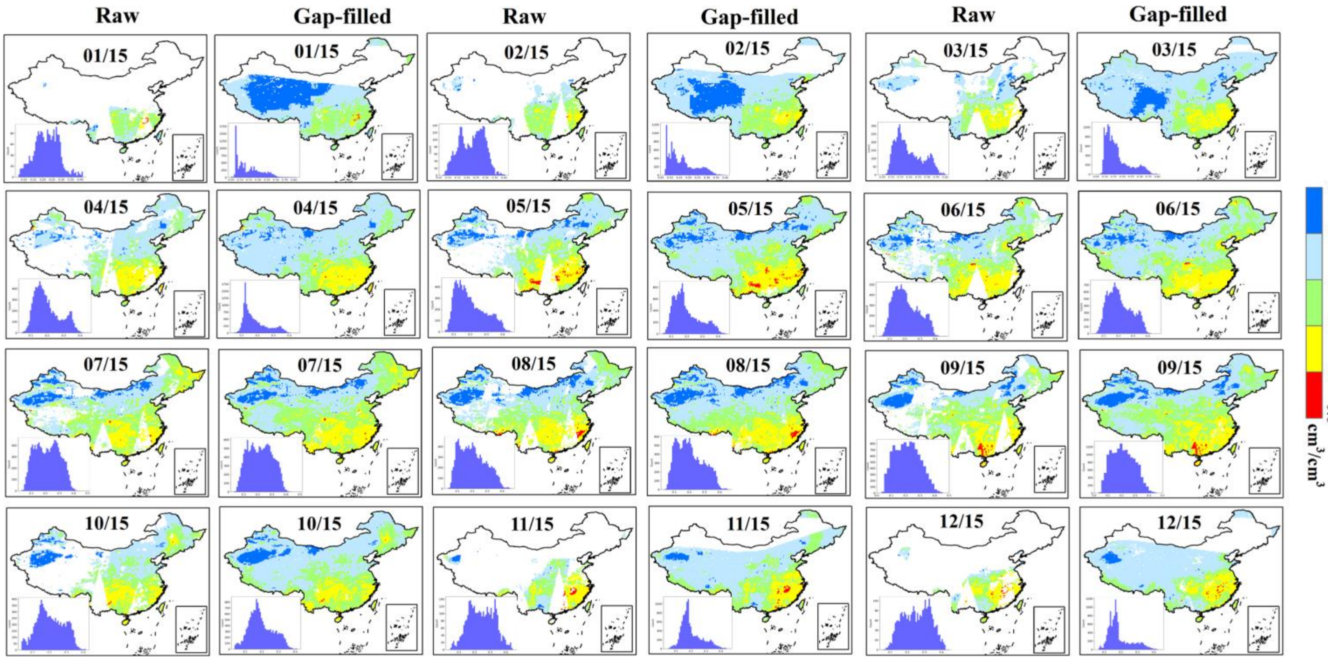


Figure 6: The spatial distributions and histogram of the raw and gap-filled CCI SM on the 15th of each month in 2009.

4.2 Accuracy validation

The proposed model is first evaluated with sparse in situ measurements from WATER and CERN. As shown in Fig. 7(a), agreement is obtained between the 1-km CCI SM-derived values and the in situ measurements, with an R^2 of 0.8. This accordance is also found between the 1-km reconstructed SM and the in situ measurements (Fig. 7(b)), with the R^2 of 0.75. High accuracy is also observed when performing evaluation with in situ measurements from national agro-meteorological stations. The R^2 value between the 1-km CCI SM-derived values and the in situ measurements is 0.81, while the R^2 value between the 1-km reconstructed SM and the in situ measurements is 0.71 (Fig. 7(c) and (d)). Inconsistency evidently remains, and noticeable overestimations are observed in the high range of SM. Additionally, the accuracy of the gap-filling products tends to be diminished by drought conditions, but this impact is limited.

We further validate the reconstructed results with the dense in situ measurements from the Maqu network. The RMSE and MAE values are 0.11 and 0.09 cm^3/cm^3 (Fig. 7(e)), respectively, for the 1-km CCI SM-derived values, and 0.12 and 0.09 cm^3/cm^3 (Fig. 7(f)), respectively, for the 1-km reconstructed SM. It means that reasonable agreement is obtained for both the CCI SM product and the gap-filled SM; however, poor performance is found in the range of low values mostly because of the extreme conditions and the fewer samples available for model regression.

The time series of average 0.25° CCI SM values and reconstructed SM over the dense grid are compared with the dense in situ observations. Both the original and the reconstructed SM match well with the in situ series, with NSE values of 0.83 and 0.85, respectively. The reconstructed SM (Fig. 7(g)) mostly describes the temporal dynamics of in situ measurements; that is, sufficiently capturing seasonal and daily variability. In addition, the rainfall events impacting the surface dynamics are observed to be well depicted in the SM temporal variations. The reconstructed SM appears to have inherited the merits of stability between April and November from the CCI SM, i.e., having comparable values during this period.

Cross-validation analysis is further performed with 2009 data to evaluate model performance. The obtained metrics (Fig. 8(a)) illustrate reasonable coincidence between the reconstructed and the original CCI SM, with a median R^2 range of 0.51 and 0.63. Better accuracy is also demonstrated by the metrics of RMSE, MAE, and ubRMSE. In particular, the median of BIAS is less than 0.01 cm^3/cm^3. Comparatively, better accuracy is achieved in the growth seasons (March–October), which can be attributed to the fact that the critical environmental factors, such as NDVI, DTR, and ERA SM, are more related to satellite-derived SM during the season of vegetation growth (Chen et al., 2014; Otkin et al., 2016).

Figure 8(b) shows the accuracy metrics for different climate regions. A pattern similar to that of the monthly means is observed, that is, acceptable accuracy occurs in most regions. No significant differences in median R^2 and BIAS are evident between the reconstructed SM of each climate region, with the bias between the maximum and minimum median R^2 and BIAS values being less than 0.09 and 0.003 cm^3/cm^3, respectively. The metrics indicate relatively poor performance (in wet regions having high specific heat capacity and low albedo. The lower amounts and high thermal entropy of the available variables (i.e., LST and albedo) in these areas can affect model capacity and stability (Wang et al., 2005). Notably, despite the relatively high RMSE, MAE, and ubRMSE values in the humid region, the R^2 value is very high (Fig. 10), which might be attributable to the high SM variability in these areas. The accuracy is lower over the regions that experience drought due to perturbations of the soil water content, but without noticeably poor performances.

The spatial distributions of the accuracy metrics in Fig. 9 further illustrate the accuracy of the proposed gap-filling model. Discrepancies are observed in some grids, but they rarely exceed 0.09 cm^3/cm^3 in absolute value. Spatially, the distribution of reconstructed SM follows a geographic gradient. The relatively low accuracies occur in areas of complex terrain in western China. For these regions, complex atmospheric conditions caused by high elevations tend to affect the simulation of surface parameters. Complex topography can result in a complicated directional anisotropy, bringing great uncertainty in modelling surface energy and water cycles (Hu et al., 2016).

The gap-filling model could be sensitive to irrigation and drought owing to the induced inhibition and water stress of vegetation. On the one hand, lower accuracy is found as expected over a considerable fraction of irrigated cropland (e.g., Northern China), which can be partly attributed to the human irrigation drain. On the other hand, focused analyses illustrate the consistency of the gap-filling SM with the in-situ measurements and the original SM under extremely dry conditions (Fig. S4), illustrating the physical plausibility of the gap-filled values for specific application.

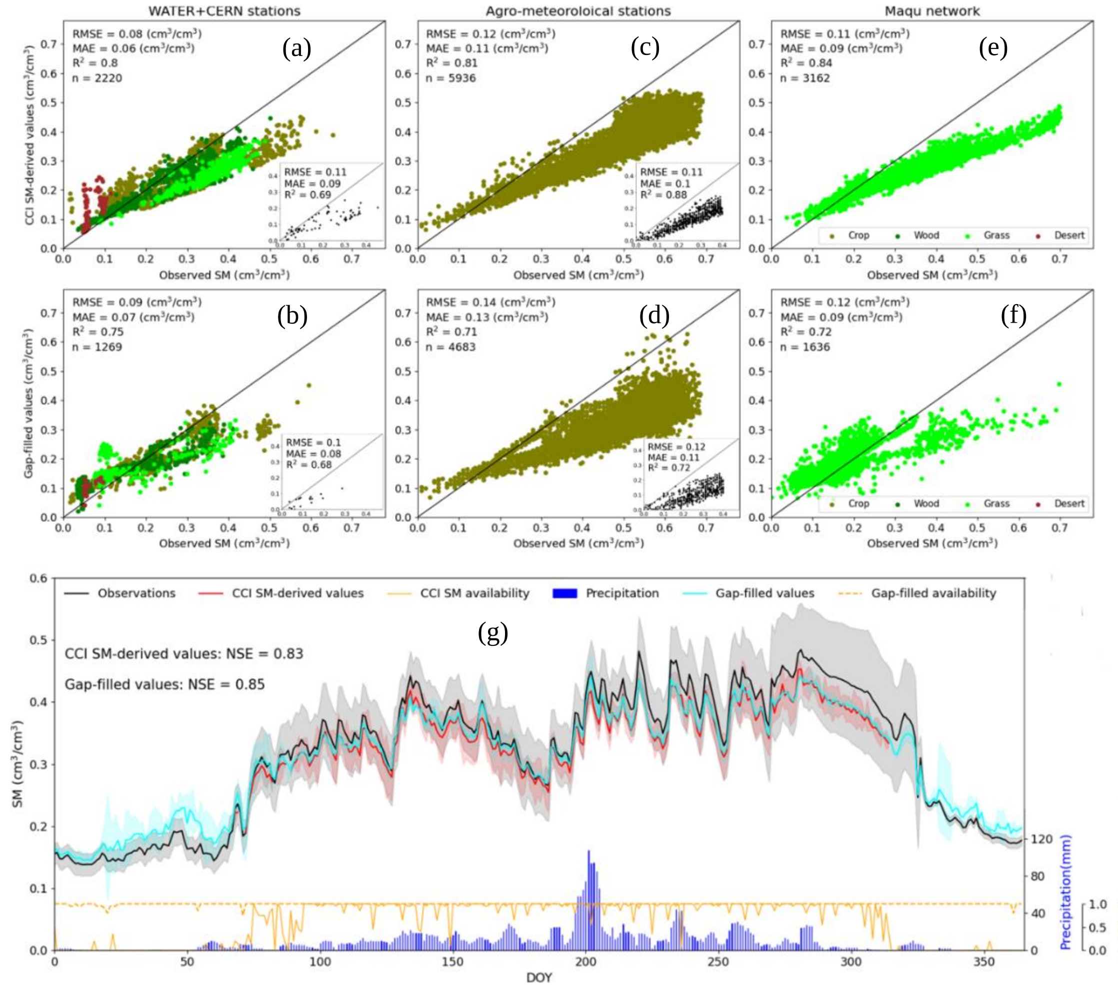


Figure 7: The evaluations of model results. (a), (c) and (e) are the scatter plots of 1-km CCI SM-derived values against field measures regarding WATER/CEERN, agro-meteorological stations, and Mauqu network, respectively, and (b), (d) and (f) are the scatter plots of 1-km gap-filled SM-derived values against field measures. The sub-figures in the upper corner of (a)-(d) are the scatter plots under extremely dry conditions. (g) are the time series of average CCI SM-derived values against site measures in the Maqu region. The shaded area in (g) denotes ±1 standard error.

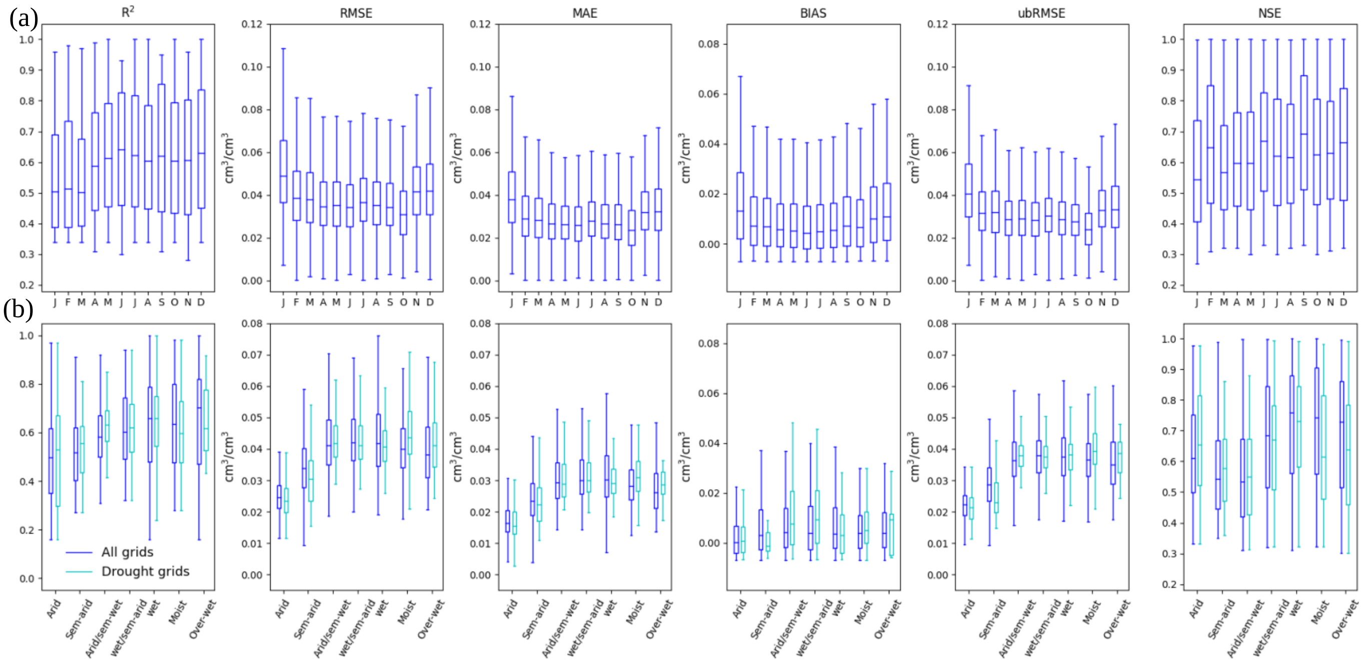


Figure 8: The accuracy metrics of 10-cross validation for R^2, RMSE, MAE, BIAS, ubRMSE and NSE: (a) is averagely obtained on a month basis, and (b) is averagely obtained for each climate region and for the drought grids.

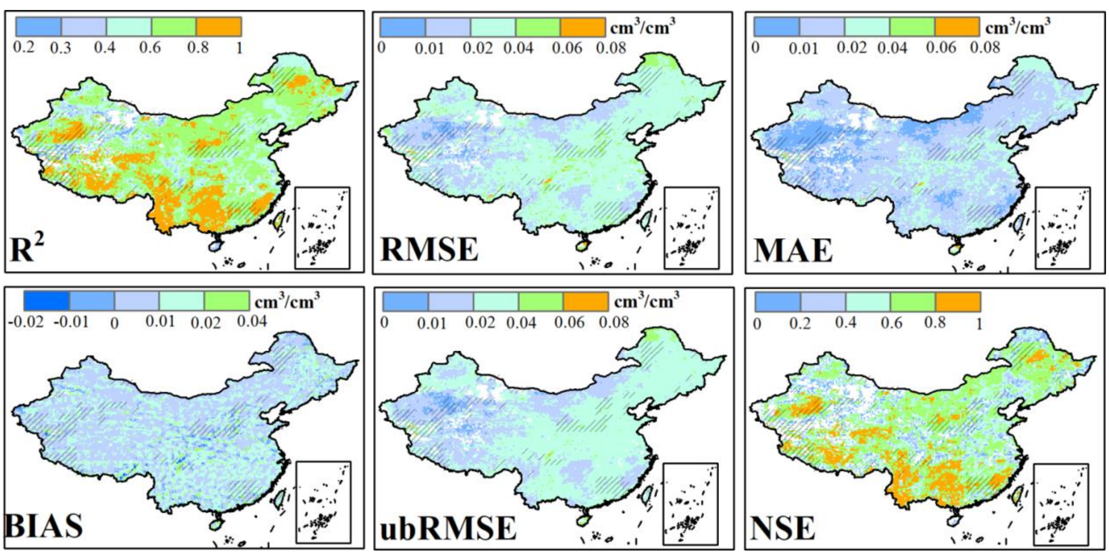


Figure 9: The spatial distributions of accuracy metrics of 10-cross validation in 2009 for R2, RMSE, MAE, BIAS, ubRMSE and NSE. The slash represents the regions impacted by drought.

4.3 Comparison analysis

The proposed method is further compared against four extensively used models, and the accuracy metrics of the five models are shown in Fig. 10. Generally, the MLR, XGB, SVM, and ANN, accompanying the RF, could potentially reconstruct the missing CCI SM pixels, indicating the stable suitability of these models and the feasibility of available variables. Moreover, the RF model demonstrates prominent performance among all the tested models, further demonstrating its capacity for reconstructing SM when integrating an effective dataset source and mining method. Our results are consistent with earlier studies that illustrated the robustness of the RF approach in simulating satellite parameters (Karbalaye Ghorbanpour et al., 2021; Zhao et al., 2018). This is attributed to the strong capacity of the RF method for coping with sparse samples, in addition to the fact that the RF does not assume a specific functional or geometric form of the model. We also check the accuracy of the models excluding the residual calibration procedure, which is an essential component of the proposed model. Results (in Fig. 10) demonstrate that accuracies are lowered by ~9% when removing the residual calibration, underscoring the importance of residual modulation in improving SM reconstruction. Moreover, better performance brought by the spatiotemporal domain strategy is also exhibited when compared with the global regression. Quantitatively, the spatiotemporal domains can improve the accuracy by ~19% in forcing the RF regression. Overall, these analyses indicate the feasibility of the proposed model by integrating the modules of the residual calibration and the spatiotemporal domain strategy.

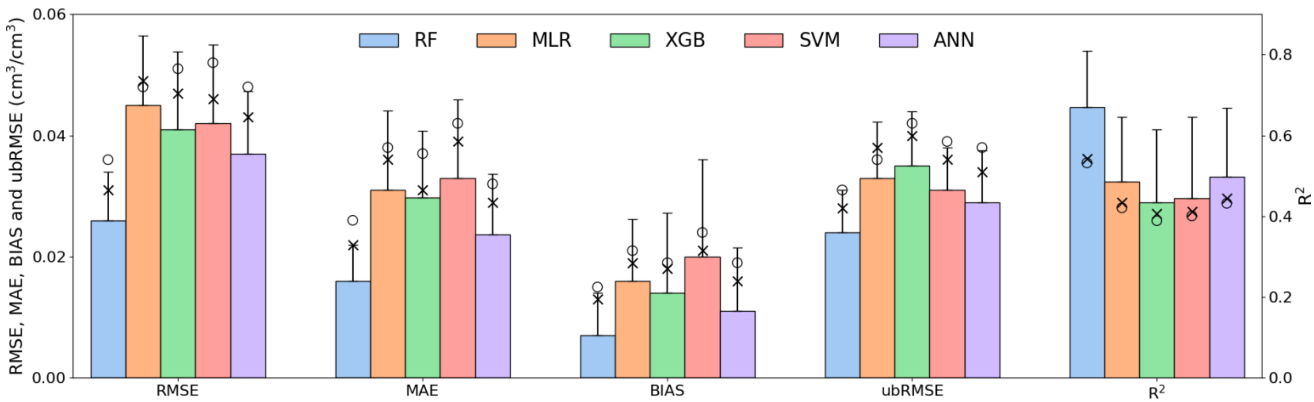


Figure 10: Comparison RF-based model with other models (i.e., MLR, XGB, SVM and ANN). Error bars denote 1σ errors. The symbol 'x' represents the accuracy metrics of models excluding the residual calibration, and the symbol 'o' represents the accuracy metrics of the models that use the global regression rather than regional regression based on the spatiotemporal window searching strategy.

4.4 Uncertainty analysis

We investigate the accuracy of the reconstruction model that excludes one participating variable. As illustrated in Fig. 11(a), the performance of the model with six variables (i.e., excluding one) is relatively low when compared with that of a model with seven variables. The strategy of removing one variable can lower the accuracy by 2.2–6.4% in terms of R^2 and by 10–30% in terms of BIAS. This diminished performance is plausible because SM is heavily related to all the selected variables.

Specifically, variability in land surface characteristics (NDVI and albedo) and atmospheric conditions (i.e., precipitation and PET) can impact SM variability. This is plausible because satellite SM retrievals represent the signals from the upper soil layer, which is directly exposed to the land and the atmosphere. Meanwhile, additional covariates mean an increase in the number of samples participating in the regression model, therefore potentially resulting in improvement of overall accuracy. We observe that the lowest accuracy occurs when DTR is excluded, underscoring the vital role of DTR in modelling SM. The importance scores produced by the RF algorithm (Zhao et al., 2019b; Ramoelo et al., 2015) (Fig. S5) also show that all selected variables substantially impact the CCI SM simulations. Specifically, DTR shows the greatest importance, mainly relating to the fact that temperature variations might influence SM fluctuation. This supports the higher model performance observed in warm seasons, during which PET, albedo, and NDVI exhibit a higher importance score. During this period, heat from the surface can be transferred to the atmosphere via ET and sensible heat conduction, thereby modifying surface SM variations (Amani et al., 2017).

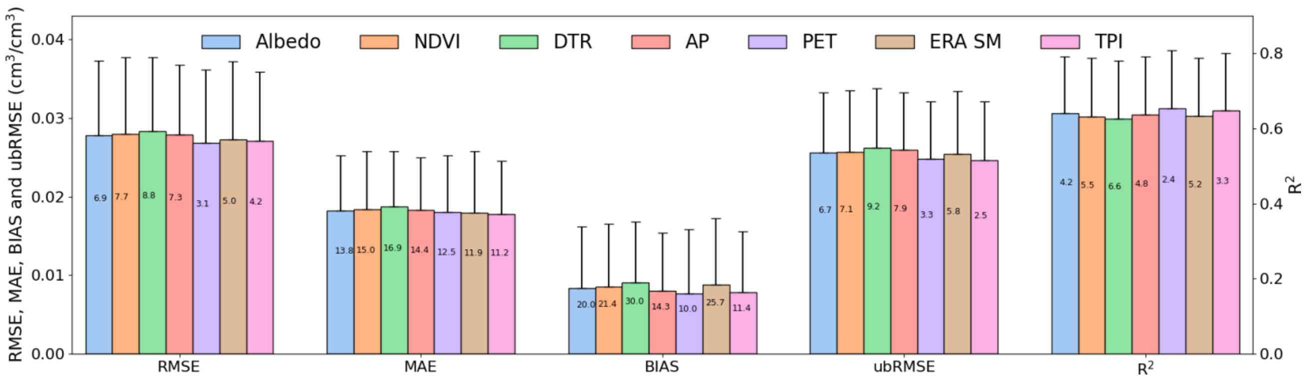


Figure 11: (a) The accuracy of the models removing one variable, i.e., using other six variables in model regression. Error bars denote 1σ errors. The text denotes the relative percentage of the decreased accuracy of the model with six variables (i.e., excluding one) in comparison with that of a model with seven variables.

We further investigate the substitution performance of other surface temperature products in reconstructing SM. Considering the bias between satellite-derived LST and modelled surface temperature, the variable correction described in section 3.1.5 is conducted to remove the systematic bias and make the simulated DTR comparable with the satellite observations. Minor reductions are found in the Pearson correlation and RF-derived importance score of three numerical model- simulated DTRs (Fig. S6) when compared with the MODIS-derived DTR, which indicates the feasibility of using each of these datasets in reconstructing SM. Reductions in model accuracy are evident when replacing the satellite-derived LST with the other three simulated sources (Fig. 12(a) and (b)). Nevertheless, the availability of reconstructed SM products is remarkably increased (by ~6–11%) owing to the all-weather coverage of the reanalysis and land surface model simulations. The surface temperature source from the numerical model dataset is suggested to be an alternative for satellite LST, which is essential on the long-term and large extended scale, especially considering their full-coverage characteristic. However, in comparison with the results obtained using the correction procedure, reduction in accuracy metrics (~4%) occurs when not considering the variable correction procedure. It emphasizes the indispensable contribution of the variable calibration procedures in reconstructing surface characteristics (Duan and Bastiaanssen, 2013; Liu et al., 2020a).

We also compare the ERA SM with two other products to evaluate its feasibility in reconstructing CCI SM. GLEAM and Noah surface SM are separately employed to replace the ERA SM while keeping other explanatory variables the same. Although the GLEAM and Noah SM-based schemes can demonstrate acceptable accuracies, they exhibit slightly inferior accuracies in comparison with the ERA SM-based schemes, probably owing to their relatively large uncertainties in depicting the surface SM dynamics across the two selected regions. Nevertheless, our study focuses on only two local regions; therefore we cannot claim that the ERA product could provide the best performance across China, and more attention should be focused on this in further work.

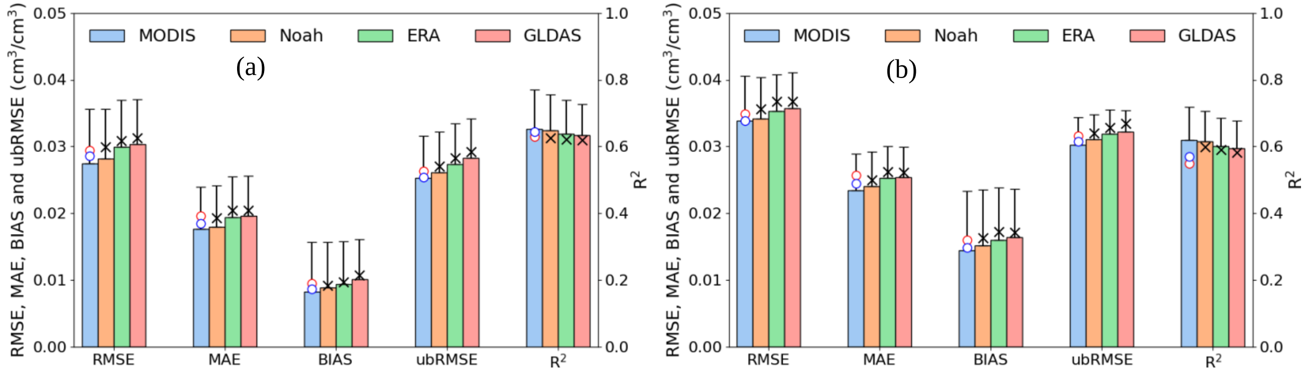


Figure 12: The metrics of models using different DTRs for (a) Northern China (NC) and (b) Southern China (SC). Error bars denote 1σ errors. The symbol 'x' represents the accuracy metrics of the models without DTR correction procedure. The symbol 'o' in red represents the accuracy metrics of the models using GLEAM SM to replace ERA SM, and the symbol 'o' in blue represents the accuracy of the models using Noah SM to replace ERA SM.

4.5 Long-term extension

The proposed gap-filling method is further extended to the long-term ECA CCI SM databases. During 2005–2015, more than 90% of contaminated pixels can be reconstructed using our model. When evaluating the pixels against in situ measurements from the dense Maqu network, we observe that the reconstructed SM during 2005—2015 has accuracy that is comparable to that in 2009 (Table 4). The average R^2 and RMSE values of the reconstructed SM are 0.73 and 0.12 cm^3/cm^3, respectively. The present results indicate that the proposed model has strong capacity for simulation of SM on the long-term scale.

The spatial distribution and the obvious differences between the gap-filled and original SM dataset can be seen in Figure 13 (a)-(c). Negative differences in SM occur in most regions, while positive differences are evident in small areas of the wet and arid regions. The dynamics and trends of SM are fundamental to assessing and quantifying ecohydrological regime. Owing to the missing satellite retrievals, the CCI SM tends to be overestimated. As shown in Fig. 13(d)-(f), the difference in valid participating SM values causes disparity in calculating the SM trend, i.e., bringing a lower SM trend in most wet regions but a higher SM trend in some dry regions when gap-filled values are introduced. Additionally, most regions with a significant trend demonstrate a lower trend in comparison with the trends of the original SM. The confidence level of the SM trend is converted from a significance level to a non-significance level for a considerable fraction of the grids. This is more pronounced in wet regions such as northeast, northwest, and southwest parts of China, which are sensitive to monsoon precipitation and ice melting. Our results are corroborated by earlier studies (Zhang et al., 2018; Gunnarsson et al., 2021) that revealed an overestimation in the trend of missing AOD and albedo when cloudy conditions prevented satellite retrievals. It means that the variations in SM trend are related to changes in the climate variables (e.g., precipitation) and land management activities (Li et al., 2018).

Table 4 Metrics for the gap-filling performance regarding Maqu network for the extended years

Year	R^2		RMSE (cm^3/cm^3)		MAE (cm^3/cm^3)		Bias (cm^3/cm^3)		ubRMSE (cm^3/cm^3)		NSE	
	CCI	gap-filled	CCI	gap-filled	CCI	gap-filled	CCI	gap-filled	CCI	gap-filled	CCI	gap-filled
2008	0.8	0.71	0.11	0.13	0.1	0.13	0.06	0.07	0.06	0.06	0.8	0.81
2010	0.82	0.73	0.1	0.11	0.09	0.11	0.05	0.06	0.06	0.05	0.81	0.83
2011	0.83	0.74	0.09	0.11	0.09	0.1	0.06	0.06	0.06	0.05	0.82	0.84
2012	0.81	0.72	0.12	0.13	0.09	0.12	0.06	0.05	0.05	0.05	0.81	0.82
2013	0.82	0.73	0.09	0.12	0.09	0.13	0.06	0.07	0.05	0.07	0.8	0.82
2014	0.85	0.74	0.09	0.11	0.08	0.09	0.06	0.08	0.05	0.06	0.83	0.85
2015	0.79	0.69	0.12	0.14	0.1	0.12	0.07	0.09	0.07	0.07	0.79	0.81

Note: NSE is from the evaluation with the time series of average 0.25° pixels while the other five metrics are from the evaluation with 1 km disaggregated values.

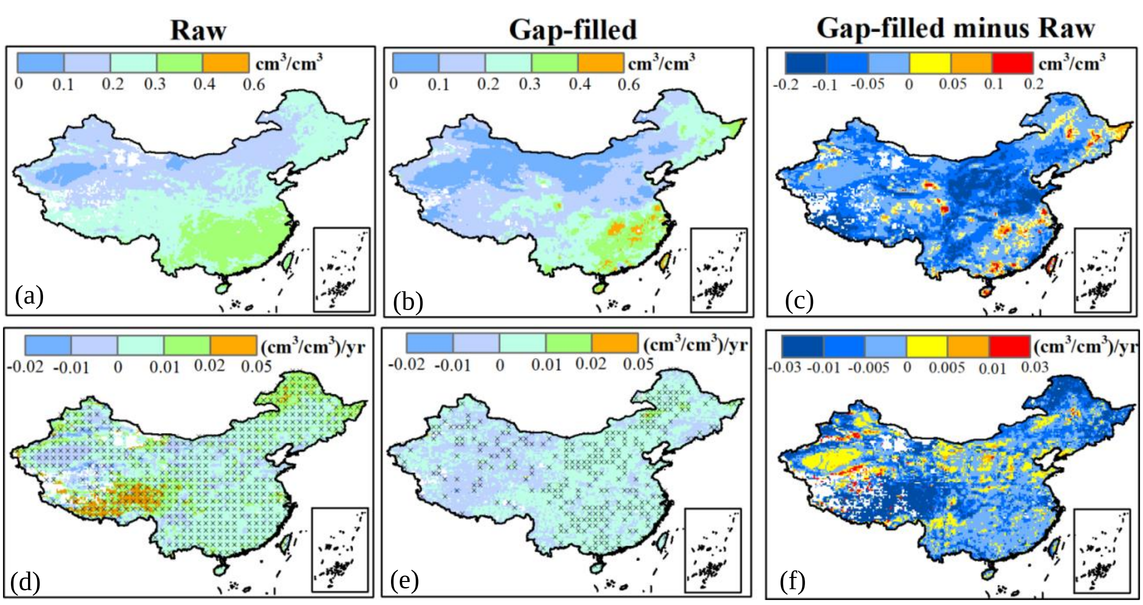


Figure 13: The implementation of the proposed model to 2005-2015. (a) and (b) are the average values of raw CCI and gap-filled SM during 2005-2015, and (c) is the difference between them. (d) and (e) are the average trends of raw CCI and gap-filled SM during 2005-2015, and (f) is the difference between them. The symbol "x" in (d) and (e) denotes the significance level under 0.05.

The biases in SM dynamics and trends are shown more pronounced for each climate region in Fig. 14(a) and 14(b). The regional averages of reconstructed SM are relatively low in comparison with those from the original CCI SM, and this pattern is clearly reflected in the trend-cycle and seasonal component (Fig. S7). The improvement of the reconstructed dataset in depicting SM trends is quantitatively manifested in Fig. 14(c)-(f), that is, the R^2 value between the trends from the original CCI SM and those from the in situ measurements is 0.28, while the R^2 value between the trends from the reconstructed CCI SM and those from the observations is increased to 0.49. Overall, an effective gap-filled model is demanded considering its capacity from depicting the dynamics and trends of SM.

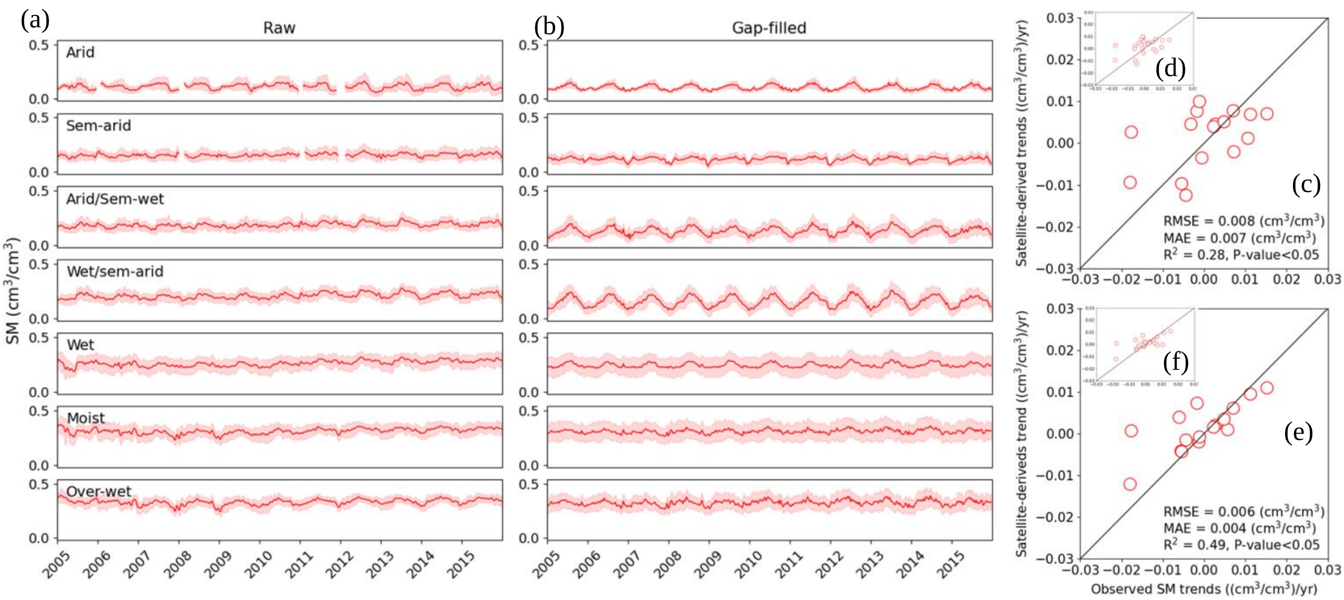


Figure 14: (a) shows the temporal patterns of raw CCI SM regarding different climate regions during 2005-2015, and (b) shows the temporal patterns of gap-filled CCI SM regarding different climate regions. The shaded area in (a) and (b) denotes ±1 standard error. (c) and (d) The scatter plot of 1-km CCI SM-derived trends against in situ measures during 2005-2014, and (c) shows the trends under significance level, while (d) shows all the trends. (e) and (f) The scatter plot of 1-km gap-filled SM-derived trends against in situ measures during 2005-2014, and (e) shows the trends under significance level, while (f) shows all the trends.

5. Conclusions and future considerations

The continuity of satellite-derived SM series is hampered by data gap problems. This study provides a novel framework for reconstructing a spatially continuous daily SM dataset by integrating the European Space Agency CCI SM and related explanatory variables. To achieve this, the random forest method taking full account of both the spatial and temporal domains is adopted. The explanatory variables filtered based on a spatiotemporal window search strategy exhibit substantial effect in driving the RF regression, resulting in efficacy improvement of ~19%. Meanwhile, model performance is enhanced by calibrating the derived residuals based on geographically weight regression and Gaussian filters. This improvement is manifested by the fact that the accuracies of gap-filling models are lowered by ~9% when removing the residual calibration procedure.

Our study illustrates the merit of identifying a sufficient number of explanatory variables from the integration of satellite observations and model-driven knowledge. This is clearly verified by the fact that the accuracy of reconstructed SM is noticeably reduced when excluding one of each of the participating variables in turn while retaining the remaining variables. The selected variables complementarily reproduce the SM dynamics in addition to capturing the spatial variations, which also implies that the nonlinear correlation between the SM and explanatory variables can be depicted on the spatiotemporal scale. In addition to the conventional variables from optical remote sensing, the essential environmental elements from model-driven knowledge are used to improve the performance of SM reconstruction. Earlier studies have suggested (Li et al., 2021a; Long et al., 2019; Shangguan et al., 2017) that reanalysis datasets and land surface model products could provide spatiotemporally continuous records, indicating the great potential of simulating land surface parameters. Here, we employ a machine learning model and a bias correction procedure for CCI SM simulation, which is expected to leverage the knowledge of the reanalysis dataset and the output from the land surface model in transfer to the CCI SM time series. The reconstructed SM achieves satisfactory accuracy over China, underscoring the importance of spatial coverage and continuity of the environmental factors from model-driven knowledge, and highlighting the need for multiple datasets to be involved in gap-filled models. We further confirm this with an uncertainty analysis showing the feasibility of using alternative data sources of DTR and SM, which is essential on the long-term scales, considering the full coverage characteristic of numerical model simulated products. Nevertheless, because numerical simulation models are generally sensitive to regional surface and climatic conditions, adoption of more effective machine learning models and bias correction strategies, as well as more representative model outputs such as CLDAS and regional numerical models, could be considered in further work (Li et al., 2022a; Li et al., 2022b).

Machine learning is recognized as a powerful tool for reconstructing contaminated values. Despite the effectiveness of the RF model for in situ SM databases, its applicability to reconstructing long-term satellite observational records, especially on the large scale, deserves careful investigation. Here, we further confirm that the RF, combined with appropriate covariates exploiting both the spatial and temporal domains, together with a model-derived residual calibration module could be a robust method for gap-filling of the CCI SM database over China. The superiority of the RF-based model in reconstructing SM is further proved by comparison with four other models. Nevertheless, more advanced machine learning strategies, such as deep neural networks (DNN) and long short-term memory (LSTM), are expected to enhance simulation accuracy. Ensemble approaches that mainly account for the scale biases among different gridded datasets are required. For example, development of a Bayesian modelling framework that can provide simulation standard error using uncertainty quantification is encouraged (Zhao et al., 2019a).

The variables forcing the proposed model are all available on the long-term scale globally. Accordingly, our framework could be extended to generate a promising long-term gap-filled SM dataset. This is critical considering that spatiotemporally continuous SM is demanded for ecological and hydrological research. Thus, the findings of our study might provide insights regarding continuous monitoring of surface water dynamics and drought, and promote further research of water resources management and climate change.

Code/Data availability

All the datasets used in this study are open to the public. The National Aeronautics and Space Administration team provides the MODIS products, SRTM DEM data and GLDAS data. The ESA CCI soil moisture dataset and ERA-5 reanalysis datasets is collected from the European Centre for Medium-range Weather Forecasts (ECMWF). The Brecht Martens, Diego Miralles and their team provides the GLEAM datasets (http://www.gleam.eu/). The China Watershed Allied Telemetry Experimental Research (WATER) project, Chinese Ecosystem Research Network (CERN), and Maqu soil moisture monitoring network provides available in situ measurements at the website http://data.tpdc.ac.cn/en/. The Chinese regional ground meteorological dataset is collected from the National Tibetan Plateau Data Center (http://data.tpdc.ac.cn).

Author contribution

Kai Liu, Xueke Li, and Shudong Wang designed the theoretical formalism. Kai Liu performed the analytic calculations. Both Shudong Wang and Hongyan Zhang contributed to the final version of the paper.

Competing interests

The contact author has declared that neither they nor their co-authors have any competing interests.

Acknowledgments

This study was jointly supported by the Natural Science Foundation of China (42141007 and 41671362), and the Inner Mongolia Autonomous Region Science and Technology Achievement Transformation Special Fund Project (2021CG0045).

References

Almendra-Martín, L., Martínez-Fernández, J., Piles, M., and González-Zamora, Á.: Comparison of gap-filling techniques applied to the CCI soil moisture database in Southern Europe, Remote Sensing of Environment, 258, 112377, https://doi.org/10.1016/j.rse.2021.112377, 2021.

Amani, M., Salehi, B., Mahdavi, S., Masjedi, A., and Dehnavi, S.: Temperature-Vegetation-soil Moisture Dryness Index (TVMDI), Remote Sensing of Environment, 197, 1-14, https://doi.org/10.1016/j.rse.2017.05.026, 2017.

Balsamo, G., Albergel, C., Beljaars, A., Boussetta, S., Brun, E., Cloke, H., Dee, D., Dutra, E., Muñoz-Sabater, J., Pappenberger, F., de Rosnay, P., Stockdale, T., and Vitart, F.: ERA-Interim/Land: a global land surface reanalysis data set, Hydrol. Earth Syst. Sci., 19, 389-407, 10.5194/hess-19-389-2015, 2015.

Belgiu, M. and Drăguţ, L.: Random forest in remote sensing: A review of applications and future directions, ISPRS Journal of Photogrammetry and Remote Sensing, 114, 24-31, https://doi.org/10.1016/j.isprsjprs.2016.01.011, 2016.

Bessenbacher, V., Gudmundsson, L., and Seneviratne, S. I.: Capturing future soil-moisture droughts from irregularly distributed ground observations, Copernicus Meetings, https://doi.org/10.5194/egusphere-egu22-8714, 2022a.

Bessenbacher, V., Seneviratne, S. I., and Gudmundsson, L.: CLIMFILL v0.9: a framework for intelligently gap filling Earth observations, Geosci. Model Dev., 15, 4569-4596, 10.5194/gmd-15-4569-2022, 2022b.

Breiman, L.: Random Forests, Machine Learning, 45, 5-32, 10.1023/A:1010933404324, 2001.

Chen, B., Xu, G., Coops, N. C., Ciais, P., Innes, J. L., Wang, G., Myneni, R. B., Wang, T., Krzyzanowski, J., Li, Q., Cao, L., and Liu, Y.: Changes in vegetation photosynthetic activity trends across the Asia–Pacific region over the last three decades, Remote Sensing of Environment, 144, 28-41, https://doi.org/10.1016/j.rse.2013.12.018, 2014.

Chen, Y., Yang, K., Qin, J., Zhao, L., Tang, W., and Han, M.: Evaluation of AMSR-E retrievals and GLDAS simulations against observations of a soil moisture network on the central Tibetan Plateau, Journal of Geophysical Research: Atmospheres, 118, 4466-4475, https://doi.org/10.1002/jgrd.50301, 2013.

Cristea, N. C., Breckheimer, I., Raleigh, M. S., HilleRisLambers, J., and Lundquist, J. D.: An evaluation of terrain-based downscaling of fractional snow covered area data sets based on LiDAR-derived snow data and orthoimagery, Water Resources Research, 53, 6802-6820, https://doi.org/10.1002/2017WR020799, 2017.

Cui, Y., Yang, X., Chen, X., Fan, W., Zeng, C., Xiong, W., and Hong, Y.: A two-step fusion framework for quality improvement of a remotely sensed soil moisture product: A case study for the ECV product over the Tibetan Plateau, Journal of Hydrology, 587, 124993, https://doi.org/10.1016/j.jhydrol.2020.124993, 2020.

Cui, Y., Zeng, C., Zhou, J., Xie, H., Wan, W., Hu, L., Xiong, W., Chen, X., Fan, W., and Hong, Y.: A spatio-temporal continuous soil moisture dataset over the Tibet Plateau from 2002 to 2015, Scientific Data, 6, 247, 10.1038/s41597-019-0228-x, 2019.

Dente, L., Vekerdy, Z., Wen, J., and Su, Z.: Maqu network for validation of satellite-derived soil moisture products, International Journal of Applied Earth Observation and Geoinformation, 17, 55-65, https://doi.org/10.1016/j.jag.2011.11.004, 2012.

Detto, M., Montaldo, N., Albertson, J. D., Mancini, M., and Katul, G.: Soil moisture and vegetation controls on evapotranspiration in a heterogeneous Mediterranean ecosystem on Sardinia, Italy, Water Resources Research, 42, https://doi.org/10.1029/2005WR004693, 2006.

Dorigo, W., Wagner, W., Albergel, C., Albrecht, F., Balsamo, G., Brocca, L., Chung, D., Ertl, M., Forkel, M., Gruber, A., Haas, E., Hamer, P. D., Hirschi, M., Ikonen, J., de Jeu, R., Kidd, R., Lahoz, W., Liu, Y. Y., Miralles, D., Mistelbauer, T., Nicolai-Shaw, N., Parinussa, R., Pratola, C., Reimer, C., van der Schalie, R., Seneviratne, S. I., Smolander, T., and Lecomte, P.: ESA CCI Soil Moisture for improved Earth system understanding: State-of-the art and future directions, Remote Sensing of Environment, 203, 185-215, https://doi.org/10.1016/j.rse.2017.07.001, 2017.

Dorigo, W. A., Gruber, A., De Jeu, R. A. M., Wagner, W., Stacke, T., Loew, A., Albergel, C., Brocca, L., Chung, D., Parinussa, R. M., and Kidd, R.: Evaluation of the ESA CCI soil moisture product using ground-based observations, Remote Sensing of Environment, 162, 380-395, https://doi.org/10.1016/j.rse.2014.07.023, 2015.

Dorigo, W. A., Wagner, W., Hohensinn, R., Hahn, S., Paulik, C., Xaver, A., Gruber, A., Drusch, M., Mecklenburg, S., van Oevelen, P., Robock, A., and Jackson, T.: The International Soil Moisture Network: a data hosting facility for global in situ soil moisture measurements, Hydrol. Earth Syst. Sci., 15, 1675-1698, 10.5194/hess-15-1675-2011, 2011.

Duan, Z. and Bastiaanssen, W. G. M.: First results from Version 7 TRMM 3B43 precipitation product in combination with a new downscaling–calibration procedure, Remote Sensing of Environment, 131, 1-13, https://doi.org/10.1016/j.rse.2012.12.002, 2013.

ElSaadani, M., Habib, E., Abdelhameed, A. M., and Bayoumi, M.: Assessment of a Spatiotemporal Deep Learning Approach for Soil Moisture Prediction and Filling the Gaps in Between Soil Moisture Observations, Frontiers in Artificial Intelligence, 4, 10.3389/frai.2021.636234, 2021.

Entekhabi, D., Njoku, E. G., Neill, P. E. O., Kellogg, K. H., Crow, W. T., Edelstein, W. N., Entin, J. K., Goodman, S. D., Jackson, T. J., Johnson, J., Kimball, J., Piepmeier, J. R., Koster, R. D., Martin, N., McDonald, K. C., Moghaddam, M., Moran, S., Reichle, R., Shi, J. C., Spencer, M. W., Thurman, S. W., Tsang, L., and Zyl, J. V.: The Soil Moisture Active Passive (SMAP) Mission, Proceedings of the IEEE, 98, 704-716, 10.1109/JPROC.2010.2043918, 2010.

Ford, T. W. and Quiring, S. M.: Comparison and application of multiple methods for temporal interpolation of daily soil moisture, International Journal of Climatology, 34, 2604-2621, https://doi.org/10.1002/joc.3862, 2014.

Fu, G., Crosbie, R. S., Barron, O., Charles, S. P., Dawes, W., Shi, X., Van Niel, T., and Li, C.: Attributing variations of temporal and spatial groundwater recharge: A statistical analysis of climatic and non-climatic factors, Journal of Hydrology, 568, 816-834, https://doi.org/10.1016/j.jhydrol.2018.11.022, 2019.

GCOS: Implementation plan for the global observing system for climate in support of the UNFCCC (2010 update), 2010.

Gruber, A., Scanlon, T., van der Schalie, R., Wagner, W., and Dorigo, W.: Evolution of the ESA CCI Soil Moisture climate data records and their underlying merging methodology, Earth Syst. Sci. Data, 11, 717-739, 10.5194/essd-11-717-2019, 2019.

Guisan, A., Weiss, S. B., and Weiss, A. D.: GLM versus CCA spatial modeling of plant species distribution, Plant Ecology, 143, 107-122, 10.1023/A:1009841519580, 1999.

Gunnarsson, A., Gardarsson, S. M., Pálsson, F., Jóhannesson, T., and Sveinsson, Ó. G. B.: Annual and inter-annual variability and trends of albedo of Icelandic glaciers, The Cryosphere, 15, 547-570, 10.5194/tc-15-547-2021, 2021.

He, J., Yang, K., Tang, W., Lu, H., Qin, J., Chen, Y., and Li, X.: The first high-resolution meteorological forcing dataset for land process studies over China, Scientific Data, 7, 25, 10.1038/s41597-020-0369-y, 2020.

Hu, L., Monaghan, A., Voogt, J. A., and Barlage, M.: A first satellite-based observational assessment of urban thermal anisotropy, Remote Sensing of Environment, 181, 111-121, https://doi.org/10.1016/j.rse.2016.03.043, 2016.

Jing, W., Zhang, P., and Zhao, X.: Reconstructing Monthly ECV Global Soil Moisture with an Improved Spatial Resolution, Water Resources Management, 32, 2523-2537, 10.1007/s11269-018-1944-2, 2018.

Karbalaye Ghorbanpour, A., Hessels, T., Moghim, S., and Afshar, A.: Comparison and assessment of spatial downscaling methods for enhancing the accuracy of satellite-based precipitation over Lake Urmia Basin, Journal of Hydrology, 596, 126055, https://doi.org/10.1016/j.jhydrol.2021.126055, 2021.

Kerr, Y. H., Waldteufel, P., Wigneron, J., Martinuzzi, J., Font, J., and Berger, M.: Soil moisture retrieval from space: the Soil Moisture and Ocean Salinity (SMOS) mission, IEEE Transactions on Geoscience and Remote Sensing, 39, 1729-1735, 10.1109/36.942551, 2001.

Leng, P., Li, Z.-L., Duan, S.-B., Gao, M.-F., and Huo, H.-Y.: A practical approach for deriving all-weather soil moisture content using combined satellite and meteorological data, ISPRS Journal of Photogrammetry and Remote Sensing, 131, 40-51, https://doi.org/10.1016/j.isprsjprs.2017.07.013, 2017.

Li, B., Liang, S., Liu, X., Ma, H., Chen, Y., Liang, T., and He, T.: Estimation of all-sky 1 km land surface temperature over the conterminous United States, Remote Sensing of Environment, 266, 112707, https://doi.org/10.1016/j.rse.2021.112707, 2021a.

Li, L., Dai, Y., Shangguan, W., Wei, N., Wei, Z., and Gupta, S.: Multistep Forecasting of Soil Moisture Using Spatiotemporal Deep Encoder–Decoder Networks, Journal of Hydrometeorology, 23, 337-350, 10.1175/jhm-d-21-0131.1, 2022a.

Li, L., Dai, Y., Shangguan, W., Wei, Z., Wei, N., and Li, Q.: Causality-Structured Deep Learning for Soil Moisture Predictions, Journal of Hydrometeorology, 10.1175/jhm-d-21-0206.1, 2022b.

Li, Q., Li, Z., Shangguan, W., Wang, X., Li, L., and Yu, F.: Improving soil moisture prediction using a novel encoder-decoder model with residual learning, Computers and Electronics in Agriculture, 195, 106816, https://doi.org/10.1016/j.compag.2022.106816, 2022c.

Li, Q., Wang, Z., Shangguan, W., Li, L., Yao, Y., and Yu, F.: Improved daily SMAP satellite soil moisture prediction over China using deep learning model with transfer learning, Journal of Hydrology, 600, 126698, https://doi.org/10.1016/j.jhydrol.2021.126698, 2021b.

Li, X., Liu, K., and Tian, J.: Variability, predictability, and uncertainty in global aerosols inferred from gap-filled satellite observations and an econometric modeling approach, Remote Sensing of Environment, 261, 112501, https://doi.org/10.1016/j.rse.2021.112501, 2021c.

Li, X., Zhang, C., Li, W., and Liu, K.: Evaluating the Use of DMSP/OLS Nighttime Light Imagery in Predicting PM2.5 Concentrations in the Northeastern United States, Remote Sensing, 9, 620, 2017.

Li, Y., Piao, S., Li, L. Z. X., Chen, A., Wang, X., Ciais, P., Huang, L., Lian, X., Peng, S., Zeng, Z., Wang, K., and Zhou, L.: Divergent hydrological response to large-scale afforestation and vegetation greening in China, Science Advances, 4, eaar4182, doi:10.1126/sciadv.aar4182, 2018.

Liu, K., Li, X., and Long, X.: Trends in groundwater changes driven by precipitation and anthropogenic activities on the southeast side of the Hu Line, Environmental Research Letters, 16, 094032, 10.1088/1748-9326/ac1ed8, 2021a.

Liu, K., Li, X., and Wang, S.: Characterizing the spatiotemporal response of runoff to impervious surface dynamics across three highly urbanized cities in southern China from 2000 to 2017, International Journal of Applied Earth Observation and Geoinformation, 100, 102331, https://doi.org/10.1016/j.jag.2021.102331, 2021b.

Liu, K., Su, H., Li, X., and Chen, S.: Development of a 250-m Downscaled Land Surface Temperature Data Set and Its Application to Improving Remotely Sensed Evapotranspiration Over Large Landscapes in Northern China, IEEE Transactions on Geoscience and Remote Sensing, 1-12, 10.1109/TGRS.2020.3037168, 2020a.

Liu, K., Wang, S., Li, X., and Wu, T.: Spatially Disaggregating Satellite Land Surface Temperature With a Nonlinear Model Across Agricultural Areas, Journal of Geophysical Research: Biogeosciences, 124, 3232-3251, https://doi.org/10.1029/2019JG005227, 2019.

Liu, Y., Yao, L., Jing, W., Di, L., Yang, J., and Li, Y.: Comparison of two satellite-based soil moisture reconstruction algorithms: A case study in the state of Oklahoma, USA, Journal of Hydrology, 590, 125406, https://doi.org/10.1016/j.jhydrol.2020.125406, 2020b.

Llamas, R. M., Guevara, M., Rorabaugh, D., Taufer, M., and Vargas, R.: Spatial Gap-Filling of ESA CCI Satellite-Derived Soil Moisture Based on Geostatistical Techniques and Multiple Regression, Remote Sensing, 12, 665, 2020.

Long, D., Bai, L., Yan, L., Zhang, C., Yang, W., Lei, H., Quan, J., Meng, X., and Shi, C.: Generation of spatially complete and daily continuous surface soil moisture of high spatial resolution, Remote Sensing of Environment, 233, 111364, https://doi.org/10.1016/j.rse.2019.111364, 2019.

Long, D., Yan, L., Bai, L., Zhang, C., Li, X., Lei, H., Yang, H., Tian, F., Zeng, C., Meng, X., and Shi, C.: Generation of MODIS-like land surface temperatures under all-weather conditions based on a data fusion approach, Remote Sensing of Environment, 246, 111863, https://doi.org/10.1016/j.rse.2020.111863, 2020.

Mao, H., Kathuria, D., Duffield, N., and Mohanty, B. P.: Gap Filling of High-Resolution Soil Moisture for SMAP/Sentinel-1: A Two-Layer Machine Learning-Based Framework, Water Resources Research, 55, 6986-7009, https://doi.org/10.1029/2019WR024902, 2019.

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.

Merlin, O., Jacob, F., Wigneron, J., Walker, J., and Chehbouni, G.: Multidimensional Disaggregation of Land Surface Temperature Using High-Resolution Red, Near-Infrared, Shortwave-Infrared, and Microwave-L Bands, IEEE Transactions on Geoscience and Remote Sensing, 50, 1864-1880, 10.1109/TGRS.2011.2169802, 2012.

Miralles, D. G., Holmes, T. R. H., De Jeu, R. A. M., Gash, J. H., Meesters, A. G. C. A., and Dolman, A. J.: Global land-surface evaporation estimated from satellite-based observations, Hydrol. Earth Syst. Sci., 15, 453-469, 10.5194/hess-15-453-2011, 2011.

Otkin, J. A., Anderson, M. C., Hain, C., Svoboda, M., Johnson, D., Mueller, R., Tadesse, T., Wardlow, B., and Brown, J.: Assessing the evolution of soil moisture and vegetation conditions during the 2012 United States flash drought, Agricultural and Forest Meteorology, 218-219, 230-242, https://doi.org/10.1016/j.agrformet.2015.12.065, 2016.

Prihodko, L., Denning, A. S., Hanan, N. P., Baker, I., and Davis, K.: Sensitivity, uncertainty and time dependence of parameters in a complex land surface model, Agricultural and Forest Meteorology, 148, 268-287, https://doi.org/10.1016/j.agrformet.2007.08.006, 2008.

Ramoelo, A., Cho, M. A., Mathieu, R., Madonsela, S., van de Kerchove, R., Kaszta, Z., and Wolff, E.: Monitoring grass nutrients and biomass as indicators of rangeland quality and quantity using random forest modelling and WorldView-2 data, International Journal of Applied Earth Observation and Geoinformation, 43, 43-54, https://doi.org/10.1016/j.jag.2014.12.010, 2015.

Reichle, R. H., Koster, R. D., De Lannoy, G. J. M., Forman, B. A., Liu, Q., Mahanama, S. P. P., and Touré, A.: Assessment and Enhancement of MERRA Land Surface Hydrology Estimates, Journal of Climate, 24, 6322-6338, 10.1175/jcli-d-10-05033.1, 2011.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat: Deep learning and process understanding for data-driven Earth system science, Nature, 566, 195-204, 10.1038/s41586-019-0912-1, 2019.

Schaake, J. C., Duan, Q., Koren, V., Mitchell, K. E., Houser, P. R., Wood, E. F., Robock, A., Lettenmaier, D. P., Lohmann, D., Cosgrove, B., Sheffield, J., Luo, L., Higgins, R. W., Pinker, R. T., and Tarpley, J. D.: An intercomparison of soil moisture fields in the North American Land Data Assimilation System (NLDAS), Journal of Geophysical Research: Atmospheres, 109, https://doi.org/10.1029/2002JD003309, 2004.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and Freitas, N. d.: Taking the Human Out of the Loop: A Review of Bayesian Optimization, Proceedings of the IEEE, 104, 148-175, 10.1109/JPROC.2015.2494218, 2016.

Shangguan, W., Hengl, T., Mendes de Jesus, J., Yuan, H., and Dai, Y.: Mapping the global depth to bedrock for land surface modeling, Journal of Advances in Modeling Earth Systems, 9, 65-88, https://doi.org/10.1002/2016MS000686, 2017.

Sismanidis, P., Bechtel, B., Keramitsoglou, I., Göttsche, F., and Kiranoudis, C. T.: Satellite-derived quantification of the diurnal and annual dynamics of land surface temperature, Remote Sensing of Environment, 265, 112642, https://doi.org/10.1016/j.rse.2021.112642, 2021.

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.

Stroud, J. R., Müller, P., and Sansó, B.: Dynamic models for spatiotemporal data, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63, 673-689, https://doi.org/10.1111/1467-9868.00305, 2001.

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.

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., and Feuston, B. P.: Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling, Journal of Chemical Information and Computer Sciences, 43, 1947-1958, 10.1021/ci034160g, 2003.

Uebbing, B., Forootan, E., Braakmann-Folgmann, A., and Kusche, J.: Inverting surface soil moisture information from satellite altimetry over arid and semi-arid regions, Remote Sensing of Environment, 196, 205-223, https://doi.org/10.1016/j.rse.2017.05.004, 2017.

van Zyl, J. J.: The Shuttle Radar Topography Mission (SRTM): a breakthrough in remote sensing of topography, Acta Astronautica, 48, 559-565, https://doi.org/10.1016/S0094-5765(01)00020-0, 2001.

Wanders, N., Karssenberg, D., de Roo, A., de Jong, S. M., and Bierkens, M. F. P.: The suitability of remotely sensed soil moisture for improving operational flood forecasting, Hydrol. Earth Syst. Sci., 18, 2343-2357, 10.5194/hess-18-2343-2014, 2014.

Wang, A., Lettenmaier, D. P., and Sheffield, J.: Soil Moisture Drought in China, 1950–2006, Journal of Climate, 24, 3257-3271, 10.1175/2011jcli3733.1, 2011.

Wang, C., Xie, Q., Gu, X., Yu, T., Meng, Q., Zhou, X., Han, L., and Zhan, Y.: Soil moisture estimation using Bayesian Maximum Entropy algorithm from FY3-B, MODIS and ASTER GDEM remote-sensing data in a maize region of HeBei province, China, International Journal of Remote Sensing, 41, 7018-7041, 10.1080/01431161.2020.1752953, 2020.

Wang, K., Wang, P., Liu, J., Sparrow, M., Haginoya, S., and Zhou, X.: Variation of surface albedo and soil thermal parameters with soil moisture content at a semi-desert site on the western Tibetan Plateau, Boundary-Layer Meteorology, 116, 117-129, 10.1007/s10546-004-7403-z, 2005.

Wei, F., Wang, S., Fu, B., Brandt, M., Pan, N., Wang, C., and Fensholt, R.: Nonlinear dynamics of fires in Africa over recent decades controlled by precipitation, Global Change Biology, 26, 4495-4505, https://doi.org/10.1111/gcb.15190, 2020.

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, X., Fu, B., Lü, Y., Sun, F., Wang, S., and Liu, M.: Comparison of Four Spatial Interpolation Methods for Estimating Soil Moisture in a Complex Terrain Catchment, PLOS ONE, 8, e54660, 10.1371/journal.pone.0054660, 2013.

Zhang, L., Liu, Y., Ren, L., Teuling, A. J., Zhang, X., Jiang, S., Yang, X., Wei, L., Zhong, F., and Zheng, L.: Reconstruction of ESA CCI satellite-derived soil moisture using an artificial neural network technology, Science of The Total Environment, 782, 146602, https://doi.org/10.1016/j.scitotenv.2021.146602, 2021a.

Zhang, Q., Yuan, Q., Li, J., Wang, Y., Sun, F., and Zhang, L.: Generating seamless global daily AMSR2 soil moisture (SGD-SM) long-term products for the years 2013–2019, Earth Syst. Sci. Data, 13, 1385-1401, 10.5194/essd-13-1385-2021, 2021b.

Zhang, R., Di, B., Luo, Y., Deng, X., Grieneisen, M. L., Wang, Z., Yao, G., and Zhan, Y.: A nonparametric approach to filling gaps in satellite-retrieved aerosol optical depth for estimating ambient PM2.5 levels, Environmental Pollution, 243, 998-1007, https://doi.org/10.1016/j.envpol.2018.09.052, 2018.

Zhang, X., Zhou, J., Liang, S., and Wang, D.: A practical reanalysis data and thermal infrared remote sensing data merging (RTM) method for reconstruction of a 1-km all-weather land surface temperature, Remote Sensing of Environment, 260, 112437, https://doi.org/10.1016/j.rse.2021.112437, 2021c.

Zhang, X., Chen, B., Zhao, H., Fan, H., and Zhu, D.: Soil Moisture Retrieval over a Semiarid Area by Means of PCA Dimensionality Reduction, Canadian Journal of Remote Sensing, 42, 136-144, 10.1080/07038992.2016.1175928, 2016.

Zhao, K., Wulder, M. A., Hu, T., Bright, R., Wu, Q., Qin, H., Li, Y., Toman, E., Mallick, B., Zhang, X., and Brown, M.: Detecting change-point, trend, and seasonality in satellite time series data to track abrupt changes and nonlinear dynamics: A Bayesian ensemble algorithm, Remote Sensing of Environment, 232, 111181, https://doi.org/10.1016/j.rse.2019.04.034, 2019a.

Zhao, W., Duan, S.-B., Li, A., and Yin, G.: A practical method for reducing terrain effect on land surface temperature using random forest regression, Remote Sensing of Environment, 221, 635-649, https://doi.org/10.1016/j.rse.2018.12.008, 2019b.

Zhao, W., Sánchez, N., Lu, H., and Li, A.: A spatial downscaling approach for the SMAP passive surface soil moisture product using random forest regression, Journal of Hydrology, 563, 1009-1024, https://doi.org/10.1016/j.jhydrol.2018.06.081, 2018.

Zhu, X., Liu, D., and Chen, J.: A new geostatistical approach for filling gaps in Landsat ETM+ SLC-off images, Remote Sensing of Environment, 124, 49-60, https://doi.org/10.1016/j.rse.2012.04.019, 2012.