Daily soil moisture mapping at 1 km resolution based on SMAP data for ~~areas affected by desertification~~desertification areas in Northern China

Pinzeng Rao[1,2], Yicheng Wang[2], Fang Wang[2*], Yang Liu[2], Xiaoya Wang[3], Zhu Wang[2]

[1]State Key Laboratory of Hydroscience and Engineering, Department of Hydraulic Engineering, Tsinghua University, Beijing 100084, China.
[2]State Key Laboratory of Simulation and Regulation of Water Cycle in River Basin, China Institute of Water Resources and Hydropower Research, Beijing 100038, China.
[3]State Key Laboratory of Remote Sensing Science, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China.
*Correspondence to: Fang Wang (657563390@qq.com)

Abstract: Land surface soil moisture (SM) plays a critical role in hydrological processes and terrestrial ecosystems in ~~areas affected by desertification~~desertification areas. Passive microwave remote sensing products such as the Soil Moisture Active Passive (SMAP) have been shown to monitor surface soil water well. However, the coarse spatial resolution and lack of full coverage of these products greatly limit their application in areas undergoing desertification. In order to overcome these limitations, a combination of multiple machine learning methods, including multiple linear regression (MLR), support vector regression (SVR), artificial neural networks (ANN), random forest (RF) and extreme gradient boosting (XGB), have been applied to downscale the 36 km SMAP SM products and produce higher spatial-resolution SM data based on related surface variables, such as vegetation index and surface temperature. ~~Areas affected by desertification~~Desertification areas in Northern China, which are ~~very~~ sensitive to SM, were selected as the study area, and the downscaled SM with a resolution of 1 km on a daily scale from 2015 to 2020 was produced. The results show a good performance compared with in situ observed SM data, with an average unbiased root mean square error value of 0.~~049~~057 m^3/m^3. In addition, their time series are also consistent with precipitation and perform better than ~~some~~ common gridded SM products. The data can be used to assess soil drought and provide a reference for reversing desertification in the study area. This dataset is freely available at https://doi.org/10.6084/M9.FIGSHARE.16430478.V5 (Rao et al., 2021).

Keywords: Soil moisture; SMAP; Multiple machine learning; Surface variables; Desertification.

1. Introduction

Surface soil moisture (SM) plays a very important role in water-energy cycle processes (Sandholt et al., 2002; De Santis et al., 2021) and is an important source of water for plants and soil microbes (Wang et al., 2007; Gu et al., 2008; Mallick et al., 2009). Large-scale areas of northern China are undergoing desertification because of scarce precipitation and insufficient SM. The accurate acquisition of SM is valuable to ecological conservation and revegetation in arid areas of Northern China.

In the past, SM data were mainly obtained through ground measurements or the assimilation of products based on land surface models such as the Global Land Data Assimilation System (GLDAS) (Fang and Lakshmi, 2014; Zawadzki and Kędzior, 2016; Liu et al., 2021). Although most accurate SM data at different soil depths can be obtained, field measurements and in situ observations are limited due to the high cost and labor intensity involved in their collection and are generally not representative of soil water status over larger areas (Rahimzadeh-Bajgiran et al., 2013; Zhao et al., 2018; Bai et al., 2019). With the development of remote sensing technologies, continuous SM estimates can be generated at regional and global scales (Peng et al., 2021). Compared to ground measurements, remote sensing products can provide good spatial and temporal coverage of SM with a relatively low cost to the user (Zeng et al., 2015; Zhao et al., 2018; Meng et al., 2020). Data assimilation SM products largely depend on the accuracy of the land surface model and the original inputsdata (Zawadzki and Kędzior, 2016). They generally have low accuracy in areas where ground measurements are scarce, which is a problem that can be overcome with remote sensing.

At present, there are many remotely sensed SM data, some of which are from microwave remote sensing satellites, including active and passive types. SM retrievals from active sensors like Synthetic Aperture Radar (SAR) are sensitive to scattering and greatly affected by the surface roughness and vegetation types (Lievens et al., 2011; Wagner et al., 2013). Unlike active sensors, passive microwave radiometers or sensors are rarely affected by scatteringUnlike active sensors, passive microwave radiometers or sensors have almost no scattering and generate very stable SM products (Abbaszadeh et al., 2019). Common passive microwave SM products are listed in Table 1 below. Some studies have compared these products and found that SMAP SM products have higher accuracy and robustness than other remotely sensed SM products (Liu et al., 2019; Wang et al., 2021).

Table 1: Information of five common passive microwave soil moisture (SM) products.

SM Datasets (Abbreviation)	Name	Production source	Resolution	Temporal Coverage	Equator Crossing Time
AMSR-E/Aqua Daily L3	Advanced Microwave Scanning Radiometer-Earth Observing System	National Aeronautics and Space Administration (NASA) National Snow and Ice Data Center Distributed Active Archive Center (NSIDC)	25 km; Daily	2002-2011	1:30 PM Ascending 1:30 AM Descending
SMOS	Soil Moisture and Ocean Salinity	European Space Agency (ESA)	25 km; Daily	2010-present	6:00 PM Ascending 6:00 AM Descending
FY3B	Fengyun-3B	National Satellite Meteorological Center	25 km; Daily	2011-present	1:40 PM Ascending 1:40 AM Descending
GCOM-W1/AMSR2	Advanced Microwave Scanning Radiometer 2	Japan Aerospace Exploration Agency (JAXA)	0.25°/0.1°; Daily	2012-present	1:30 PM Ascending 1:30 AM Descending
SMAP	Soil Moisture Active Passive	National Aeronautics and Space Administration (NASA)	36 km; Daily	2015-present	6:00 PM Ascending 6:00 AM Descending

Passive microwave SM products have been applied at watershed and national scale (Fang and Lakshmi, 2014; Meng et al., 2020). However, due to their coarse spatial resolution, microwave SM products have limited applicability to small-scale areas. Compared to microwave sensors, optical satellites such as MODIS and Landsat have a finer spatial resolution. Some observations generated from optical satellites provide good information about SM, such as vegetation index (VI) (usually Normalized Difference Vegetation Index (NDVI) or Enhanced Vegetation Index (EVI)) and land surface temperature (LST) (Wang et al., 2007; Sun et al., 2012). Many experiments have tried to use these two parameters from optical remote sensing to retrieve surface SM (Mallick et al., 2009; Fang et al., 2013). Based on the LST and VI triangle space, Sandholt et al. (2002) proposed the temperature vegetation dryness index (TVDI) and used it to assess the SM status. Relative SM indicators can be calculated using optical remote sensing data, Despite their higher resolution, however, reliable ground measurements or other data are still required to obtain the true value of SM. optical remote sensing data do not allow to directly retrieve true SM.

Some studies have tried to use surface variables from optical observations to improve the spatial resolution of passive remotely sensed SM products (Peng et al., 2017). Zhao et al. (2017) used the triangle method and Landsat satellite observations to disaggregate coarse-resolution SM data. Some studies have also shown that polynomial regression is effective in SM and optical observations (Zhao and Li, 2013; Piles et al., 2016). However, these methods have some shortcomings in representing the nonlinear relationship between SM and other surface variables (Zhao et al., 2018; Hu et al., 2020). Machine learning methods can be applied to show the nonlinear relationships between SM and surface variables. Random forest (RF) and artificial neural network (ANN) have been widely used in previous studies due to their high generalization ability and robustness (Yao et al., 2017; Liu et al., 2020; Demarchi et al., 2020; Chen et al., 2021). Chen et al. (2021) developed the global surface SM dataset covering 2003–2018 at 0.1° resolution with neural networks and some related variables. Im et al. (2016) used machine learning approaches (RF, boosted regression trees, and Cubist) to downscale AMSR-E SM data in South Korea and Australia and found RF to be superior to the other downscaling methods. Although these machine learning methods perform well in constructing nonlinear regression models, there are still some shortcomings. For example, neural networks are prone to overfitting when the sampling is inefficient there are inefficient samples (Piotrowski and Napiorkowski, 2013) or variables that are weakly correlated with the dependent variable (Elshorbagy and Parasuraman, 2008; Ågren et al., 2021). Since the RF algorithm uses random sampling with replacement, its simulation results will not exceed the range of training set and tend to ignore some extreme values when used as a regression model (Belgiu and Drăguţ, 2016). Also, it does not perform well when learning from an extremely imbalanced training data (Lin et al., 2021). –Extreme gradient boosting (XGB), as a new ensemble learning method (Chen and Guestrin, 2016), performs well in some fields (Wang et al., 2020; Fan et al., 2021; Ma et al., 2021), but it has rarely been used for soil moisture downscaling. Compared with methods such as RF, the XGB algorithm adopts the boosting weighted sampling method, which can reduce the impact of imbalanced data and better simulate the extreme values existing in the samples (Chen and Guestrin, 2016). The coarse-resolution remote sensed SM (>10 km) itself has ignored some maxima or minima with relatively finer-grid SMs, so a method that better simulates extreme values will obviously have certain theoretical advantages.

85   The selection of feature variables is critical for regression models. In addition to LST and VI mentioned above, variables

86 such as terrain and soil conditions also have a significant impact on SM. Abbaszadeh et al. (2019) downscaled SMAP

87 radiometer SM products over the continental United States using MODIS products (including NDVI and LST), precipitation

88 and topographic data, and also evaluated the influence of soil texture on SM. Zhao et al. (2018) added additional surface

89 variables, such as ~~LST~~ leaf area index (LAI), normalized difference water index (NDWI), surface albedo and the solar zenith

90 angle. Hu et al. (2020) added the normalized shortwave-infrared difference bare soil moisture index (NSDSI), horizontally

91 polarized Brightness Temperature (TBh) and vertically polarized Brightness Temperature (TBv) to the regression model. In

92 general, all these variables can be classified into vegetation, temperature, soil wetness, topography, and soil factors and sensors

93 conditions.

94   In recent years, the Chinese government has carried out afforestation activities in order to reverse desertification in the

95 North. Considering the role of SM in terrestrial ecosystems~~the ecological environment~~, it is urgent to obtain accurate SM with

96 high temporal and spatial resolution. This study aims to downscale SMAP SM products by constructing a nonlinear relationship

97 between SM and related surface variables by means of multiple machine learning methods and generate SM products with

98 higher temporal and spatial resolution in ~~areas affected by desertification~~desertification areas. The in situ observed SM data

99 from the Maqu Monitoring Network and Babao Monitoring Network, several gridded SM products, and precipitation and

100 temperature data from ~~131~~ meteorological stations were used for validation and analysis.

101 **2. Materials and methodology**

102 **2.1 Study area**

103   Northern China is mostly arid with an annual precipitation of generally less than ~~400~~ 600 mm. ~~The region belongs to the~~

104 ~~temperate continental monsoon climate and~~, and is subject to large-scale desertification. The desert areas of Northern China

105 are susceptible to climate and hydrological changes and have fragile ecosystems. Soil water is a key parameter in land-

106 atmosphere interactions (Ma et al., 2019),~~of the water-vapor-ecosystem,~~ and its change greatly affects the survival of

107 vegetation and agricultural production in ~~areas affected by desertification~~desertification areas. The studied area whose

108 boundaries are provided by government departments used for this study covers 3.36 million km^2, encompassing seven

109 provinces (Fig. 1). The precipitation in the study area decreases gradually from southeast to northwest, and belongs to the

110 temperate continental climate (Fig. 1). ~~T~~he terrain is complex, and the average elevation is approximately 1900 m, ranging

111 from -192 m to 7439 m.

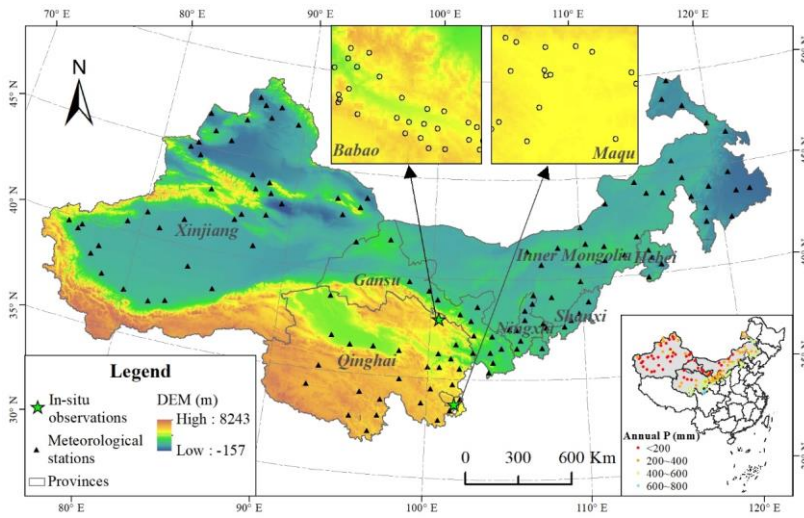



Figure 1: Location of the study area.

2.2 Observations for the production of soil moisture data

2.2.1 SMAP SM data

The SMAP satellite was launched on January 31, 2015. Its mission consists of an L-band radar and radiometer instrument suite, which provides global measurements and monitoring of SM in the top 5 cm of soil. The Level-3 products are daily composites of the Level-2 products and are the most commonly used for applications. The Level-3 products are available in three spatial resolutions: 36 km passive, 9 km active-passive, and 3 km active (O'Neill et al., 2010). Following the malfunctioning of its radar in 2015, SMAP radar data were replaced with those of Sentinel-1, limiting the application of active and active-passive products.

The SMAP Level-3 passive daily SM product (L3_SM_P, Version 6) with a grid resolution of 36 km has been produced since March 31, 2015. Zeng et al. (2015) showed that most of remotely sensed SM products were slightly better during daytime than during nighttime, and the same conclusion for the SMAP SM product was confirmed by Zhao et al. (2018). Therefore, the SMAP Level-3 SM product with the descending overpass time of 6:00 AM was used in this study. The data were downloaded from NASA Earthdata (https://search.earthdata.nasa.gov).

2.2.2 MODIS products

128   MODIS provides continuous time-series predictors for important parameters, such as vegetation index and surface

129   temperature. This paper used MODIS products MOD09A1, MOD11A1, MOD13A2, MOD15A2H and MCD43D58 (Table 2).

130   The 1-km daily LST was provided by MOD11A1, and the 1-km 16-day EVI and NDVI was provided by MOD13A2.

131   MOD15A2H provided 8-day Leaf Area Index (LAI) with a spatial resolution of 500 m. MCD43D58 provided daily albedo

132   data with a spatial resolution of 30 arc second (~1 000 m). Some soil wetness related indexes, including NDWI, NSDSI, and

133   Land Surface Water Index (LSWI), were produced by MOD09A1. Their formulas are:The soil wetness related indexes,

134   including NDWI, LSWI and NSDSI, were produced using bands of the MOD09A1 product. Their formulas are:

$$NDWI = (B_4 - B_2)/(B_4 + B_2) \qquad \underline{\qquad} \qquad (1)$$

$$LSWI = (B_2 - B_6)/(B_2 + B_6) \qquad \underline{\qquad} \qquad (2)$$

$$NSDSI = (B_6 - B_7)/B_6 \qquad \underline{\qquad} \qquad (3)$$

138   where B_2, B_4, B_6 and B_7 represent the MOD09A1 surface reflectance of the 2nd, 4th, 6th and 7th bands, respectively.

139   These MODIS products are available from NASA Earthdata (https://search.earthdata.nasa.gov)., Aand all data were

140   obtained from 2015 to 2020 and processed to a spatial resolution of 1,000 meters.-

141   **2.2.3 Topographic data**

142   Topographic factors are strongly related to SM, including elevation, slope and aspect (Bai et al., 2019; De Santis et al.,

143   2021). The Shuttle Radar Topography Mission (SRTM) digital elevation model (DEM) at 3 arc second resolution (~100 m),

144   version   3,   obtained   from   the   Land   Processes   Distributed   Active   Archive   Center   (LP

145   DAAC)(https://e4ftl01.cr.usgs.gov/MEASURES/SRTMGL3.003/), was used as elevation. Slope and aspect were generated

146   based on the DEM.

147   Topographic factors are strongly related to SM, including elevation, slope and aspect. The Shuttle Radar Topography

148   Mission (SRTM) digital elevation model (DEM) was used as elevation. Slope and aspect can be generated based on the DEM.

149   These data were obtained from the Geospatial Data Cloud (http://www.gscloud.cn/), where slope and aspect have been

150   processed and provided directly.

151   **2.2.4 Soil texture data**

152   Soil texture, the proportions of sand, silt and clay particles, controls the water holding capacity of the soil. The soil data

153   at 1,000 m resolution, including the proportions of sand, silt and clay, used for this study used the China Soil Characteristics

154   Dataset (CSCD) (Shangguan et al., 2012), obtained from National Tibetan Plateau Data Center (http://westdc.westgis.ac.cn/).

155   **2.2.5 In Situ SM observations**

156   The in situ SM measurements were collected from the data provided by the Maqu Monitoring Network (Zhang et al.,

157   2020) and the Babao Monitoring Network (Kang et al., 2017). The Maqu Monitoring Network covers 26 sites and provides

158   SM for the surface layer (0-5 cm) at 15-minute intervals from 2009 to 2019; 19 of the available sites which have data after

159 2015 were used in this study (Fig. 1). The Babao Monitoring Network covers 40 sites and provides hourly SM for the surface

160 layer (4 cm, 10 cm and 20 cm) from 2013 to 2017; 29 of the available sites have data after 2015 and their observations of the

161 first layer (4 cm) were used in this study (Fig. 1). To compare with the simulated results, they were all processed into daily

162 time series.

163 **2.2.6 Precipitation and temperature data**

164 The daily precipitation and temperature dataThe precipitation data were acquired from 131 meteorological stations from

165 the China Meteorological Data Service Centre (http://data.cma.cn). The spatial locations of these meteorological stations are

166 shown in Fig. 1. The average annual precipitation of most stations from 2015 to 2020 is less than 600 mm, and gradually

167 decreases from northwest to southeast (Fig. 1).

168 **Table 2: Main predictors used in the study and corresponding datasets**

Datasets	Predictors	Original sSpatial resolution	Temporal resolution	Number of granules (Years×tiles)
SMAP	SM	~36 km	Daily	2064
MOD11A1	LST	1 km	Daily	17460
MOD13A2	NDVI; EVI	1 km	16-day	1104
MOD15A2H	LAI; FAPAR	500 m	8-day	2208
MOD09A1	NDWI; LSWI; NSDSI	500 m	8-day	2208
MCD43D58	Albedo	30 Arcsecarc second (~1 km)	Daily	2192
SRTM	DEM; Slope; Aspect	3 arc second (~100 m)90 m	-	32
CSCD	Sand; Silt; Clay	1 km	-	1

169 **2.2.7 Other gridded SM datasets**

170 Some other gridded SM data were used to compare the simulation results (Table 3). The SMAP Level-2 product

171 (L2_SM_SP) merges SMAP radiometer and processed Sentinel-1A/1B SAR observations. It is available at 3 km and 1 km

172 resolutions. The Global Change Observation Mission Water (GCOM-W1) AMSR2 product is produced by the Japan

173 Aerospace Exploration Agency (JAXA), and SM data at a 0.1° spatial resolution were selected for this study. The Copernicus

174 Climate Change Service (C3S) produces a global SM gridded dataset from 1978 to present from satellite sensors such as SMOS,

175 AMSR2 and SMAP. It has a spatial resolution of 0.25 degrees and offers three types of products: active, passive and combined.

176 The combined product that we used in this study is generated by merging the active and passive products. The fifth generation

177 reanalysis dataset (ERA5) produced by European Centre for Medium-Range Weather Forecasts (ECMWF)ECMWF reanalysis

178 dataset (ERA5) provides several variables including volumetric soil water over several decades. In the dataset, the soil is

179 divided into four layers and the depth of the top layer is 0-7 cm. In this study, we downloaded the hourly volumetric soil water

180 data of the top layer and processed them as daily averages. Famine Early Warning Systems Network (FEWS NET) Land Data

181 Assimilation System (FLDAS) provides daily SM at a 0.01° spatial resolution over the Central Asia region (30-100° E, 21-56°

182 N), which covers part of our study area. The product consists of four layers of SM, and the SM at the top layer (0-10 cm) was

183 selected for this study.

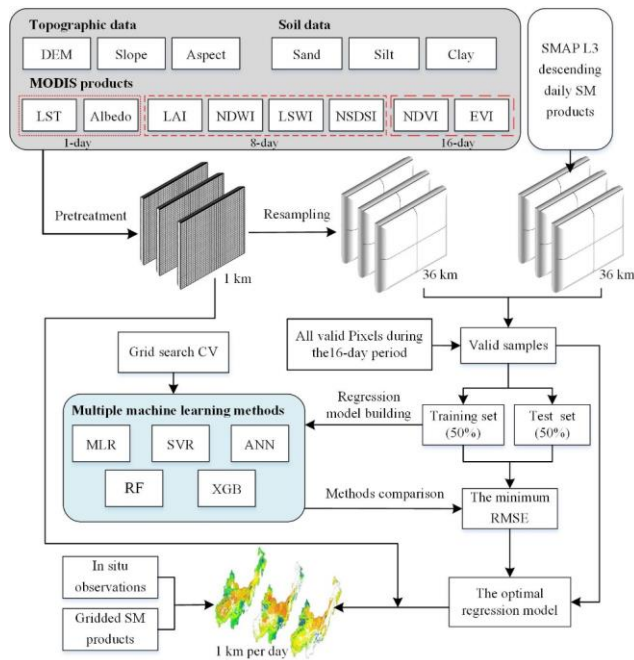
184 **Table 3: The gridded SM products used in this study**

Institution	Name	Soil layers	TYPES	Temporal resolution	Grid spacing	Data link
NASA	SMAP/ Sentinel-1 (L2_SM_SP)	One layer (0-5 cm)	Active microwave	1-2 days	1/3 km	https://cmr.earthdata.nasa.gov/search/concepts/C1931663473-NSIDC_ECS.html
JAXA	GCOM-W1/AMSR2	One layer (~)	Passive microwave	Daily	0.1°/0.25° (~11 km/28 km)	https://gportal.jaxa.jp/gpr/
ECMWF	C3S	One layer (~)	Passive, active and combined	Daily	0.25° (~28 km)	https://cds.climate.copernicus.eu/cdsapp#!/dataset/satellite-soil-moisture
ECMWF	ERA5	Four layers (0-7 cm, 7-28 cm, 28-100 cm,100-289 cm)	Reanalysis	Hourly	0.1°(~11 km)	https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land
NASA	FLDAS	Four layers (0-10 cm, 10-40 cm, 40-100 cm,100-200 cm)	Reanalysis	Daily	0.01°(~1.1 km)	https://cmr.earthdata.nasa.gov/search/concepts/C2020764153-GES_DISC.html

185 **2.3 Downscaling approach based on multi-machine learning**

186   According to the selected variable indicators (mainly including topographic data, soil data and some MODIS products)

187 and machine learning methods, we constructed a framework to downscale SMAP SM based on multiple machine learning

188 methods (Fig. 2).

190   **Figure 2: Schematic of the SMAP soil moisture downscaling framework**

2.3.1 Machine learning methods

Machine learning methods are widely used in regression and classification. We selected machine learning methods that are currently widely used to build regression models for SM and its related variables. We studied five methods: Multiple linear regression (MLR), support vector regression (SVR), artificial neural networks (ANN), random forest (RF) and extreme gradient boosting (XGB). MLR and SVR have been widely used as regression methods in the past (Yu et al., 2012; Achieng, 2019; Wang et al., 2019). ANN is currently one of the most popular machine learning methods and is used in many fields, including remote sensing of soil moisture inversion (Del Frate et al., 2003; Elshorbagy and Parasuraman, 2008; Yao et al., 2017; Chen et al., 2021).

RF and XGB are tree based ensemble algorithms, which have prediction accuracy and good generalization ability, and are not prone to overfitting (Rao et al., 2018; Abbaszadeh et al., 2019). RF is a multiple-tree algorithm improved by ~~Bootstrap~~ bootstrap to reduce decision tree bias in determining the splits (Mohana et al., 2021). Many studies have used RF to build regression models of remotely sensed SM and related variables, and almost all achieved better results compared to other

regression methods (Zhao et al., 2018; Qu et al., 2019; Hu et al., 2020). In contrast, the application of XGB, which applies a regularized gradient boosting framework, is still very limited. However, XGB has prominent~~incomparable~~ advantages in generalization performance and accuracy (Wang et al., 2020).

The XGB algorithm is a boosting-type ensemble of multiple CART decision trees (Chen and Guestrin, 2016). The predicted result of the boosting-type tree ensemble model can be expressed as follows:

$$\hat{y_i} = \emptyset(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in F \quad\quad\quad (4)$$

where F is the space of regression tree, K is the total number of trees, which means the model uses K additive functions, $f_k(x_i)$ is the weighted score of the k-th tree on i-th input data (x_i).

XGB adopts a regularized learning objective to optimize the simulation results.

$$Obj(\emptyset) = \sum_{i=1}^{N} l(y_i, \hat{y_i}) + \sum_{k=1}^{K} \Omega(f_k) \qu\quad\quad (5)$$

where l is the loss function, N is the total number of input, Ω is the regularization term to penalize the model complexity and prevent overfitting.

Compared with RF and other some methods, XGB has significantly faster calculation speed (Fan et al., 2018; Shi et al., 2021). Some studies have shown that XGB is a better regression and classification algorithm than RF and other machine learning methods (Ågren et al., 2021; Fan et al., 2021).

2.3.2 ~~Downscaling process~~ **The construction of 16-day regression model**

The downscaling process is shown in Fig. 2. First, all data need to be preprocessed. Daily LST data are likely to be affected by the cloud, so we performed quality control to MOD11A1 products using its quality control (QC) band and choose high-quality cloud-free pixels. All selected variables, including LST, Albedo, LAI, NDWI, LSWI, NSDSI, NDVI, EVI, DEM, slope, aspect, sand, silt and clay, were aggregated into a resolution of 1 km with a geotiff format. These variables were further resampled to the spatial resolution of the SMAP SM data (36 km) using the nearest neighbor interpolation method.

Second, valid samples were obtained and split. Since it is severely affected by noise (such as clouds), MOD11A1 only provides daily valid clear-sky LST values onto grids. In addition, each SMAP image has a narrow coverage and provides only a small number of valid pixels per day. It means that there may be few or no valid samples if only the data of a certain day are selected to build the regression. The variables from MOD13A2 and MOD15A2H are the best composite within 16 days and 8 days, respectively. To overcome the limitation, we chose to build regression models within 16 day periods (the lowest temporal resolution from these dynamic variables). All valid data (including training and test sets) within 16 days were used as the samples in the regression model. For instance, for NDVI and EVI on January 1, 2020, which are composite results from January 1 to January 15, the valid data during the period were used as samples. The number of valid samples for surface variables and SMAP SM for each period in 2015-2020 is shown in Fig. 3. The day of year (DOY) is used to represent the corresponding period. Since limited available SMAP SM grid data, there may be few valid samples we can obtain during cold seasons. The valid samples for each period were divided into training and test sets, each accounting for 50% of the total number of samples. In this study, stratified random sampling based on sampling date during the 16-day period was employed to split the training

带格式的：字体颜色：文字 1

236 and test sets. Moreover, to avoid excessively inconsistent training and test sets, the Kolmogorov-Smirnov (KS) test is adopted
237 to test the distribution consistency of them (Kovalev and Utkin, 2020). If the p-value of the KS test result is less than or equal
238 to 0.05, stratified random sampling is performed again, and until the requirements are met.

239 The downscaling process is shown in Fig. 2. First, due to the difference in spatial resolution and data format, all required
240 data were pre-processed. All selected variables, including LST, Albedo, LAI, NDWI, LSWI, NSDSI, NDVI, EVI, DEM, slope,
241 aspect, sand, silt and clay, were aggregated into a resolution of 1 km with a geotiff format. These variables were further
242 resampled to the spatial resolution of the SMAP SM data (36 km) using the nearest neighbor interpolation method. The
243 regression model was then defined according to the selected machine learning method:

$$SM = f(LST, Albedo, LAI, NDWI, LSWI, NSDSI, NDVI, EVI,$$
$$DEM, slope, aspect, sand, silt \ and \ clay) \tag{4}$$

246 where f represents the regression function of the machine learning method (MLR, SVR, ANN, RF or XGB).

247 Then, the regression model based on multiple machine learning was built. The MODIS products and SMAP SM data
248 have different temporal resolutions. Since it is severely affected by noise (such as clouds), MOD11A1 only provides daily
249 valid clear-sky LST values onto grids. The variables from MOD13A2 and MOD15A2H are the best composite within 16 days
250 and 8 days, respectively. In addition, because each SMAP image has a narrow coverage, there may be few or no valid samples
251 if only the data of a certain day are selected to build the regression. To overcome the limitation, we chose to build regression
252 models within 16-day periods (the lowest temporal resolution from these dynamic variables). All valid data (including training
253 and test datasets) within 16 days were used as the samples in the regression model. For instance, for NDVI and EVI on January
254 1, 2020, which are composite results from January 1 to January 15, the valid data during the period were used as samples. The
255 number of valid samples for surface variables and SMAP SM for each period in 2015-2020 is shown in Fig. 3. The day of year
256 (DOY) is used to represent the corresponding period. Since there are fewer available SMAP SM grid data in the cold season,
257 there may be few valid samples we can obtain for the period. Considering the number of samples is critical to the accuracy of
258 the regression model, we only selected periods with more than 100 samples to build the model and DOY of 2016017, 2018017,
259 2018353, 2019001 and 2019177 were excluded. The valid samples were divided into a training set and a test set, each
260 accounting for 50% of the total number of samples. Then, we used these samples and multiple machine learning methods
261 (MLR, SVR, ANN, RF and XGB) to build a regression model for each 16-day period.

262

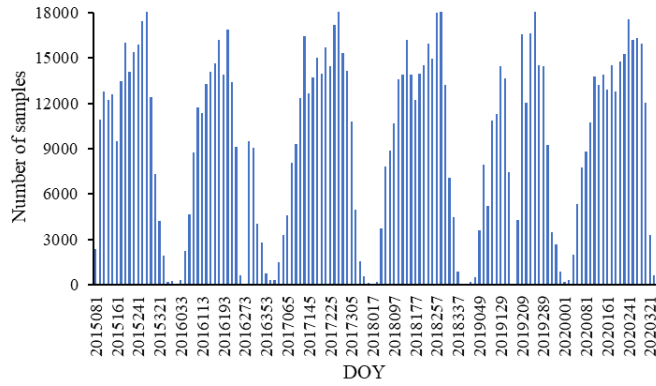



Figure 3: The number of valid samples for a 16-day period in 2015-2020. DOY is the day of year, the same below.

263

Third, the regression model was determined based on training and test sets. Considering the number of samples is critical to the accuracy of the regression model, we only selected periods with more than 100 samples to build the model and DOY of 2016017, 2018017, 2018353, 2019001 and 2019177 were excluded. Then, we used the training set and multiple machine learning methods (MLR, SVR, ANN, RF and XGB) to build a regression model for each 16-day period. The regression model was then defined according to the selected machine learning method:

$$SM = f(LST, Albedo, LAI, NDWI, LSWI, NSDSI, NDVI, EVI,$$
$$DEM, slope, aspect, sand, silt \ and \ clay) \hspace{2cm} (6)$$

where f represents the regression function of the machine learning method (MLR, SVR, ANN, RF or XGB).

Finally, hyperparameter turning and the selection of the optimal model. Hyperparameters are critical for some machine learning methods (Klein et al., 2017; Khan et al., 2020; Sun et al., 2021). In this study, the key hyperparameters of SVR, ANN, RF and XGB are tuned based on grid search cross-validation (CV). All models are evaluated based on the correlation coefficient (R) and the root mean square error (RMSE). They are calculated as:

$$R = \frac{Cov(SM_I, SM_P)}{\sqrt{Var(SM_I)Var(SM_P)}} \hspace{2cm} (7)$$

$$RMSE = \sqrt{\frac{1}{n}(SM_P - SM_I)^2} \hspace{2cm} (8)$$

where SM_I is the SMAP SM, SM_P is the corresponding SM predicted by the regression model, *Cov* represents the covariance function, *Var* is the variance, and n is the number of valid samples for SM_I or SM_P.

The RMSE is used as the evaluation metric for hyperparameter turning. The tuning results of hyperparameters are shown in Tables S2 and S3. According to the optimal hyperparameter, the corresponding model can be constructed.

2.3.3 Prediction of 1-km daily SM product

The accuracy of the five regression models is compared using the average RMSE of training and test sets. This average RMSE can be expressed as:

$$\overline{RMSE} = \frac{RMSE_{Training} + RMSE_{Test}}{2} \tag{9}$$

where $RMSE_{Training}$ and $RMSE_{Test}$ are the RMSE of training and test sets for these models, respectively.

The regression model with the smallest $\overline{RMSE}$ was selected as the optimal model. Furthermore, we used the selected optimal model and these surface variables with a resolution of 1 km within 16 days to simulate daily SM at 1 km resolution on the corresponding date. Taking 16 days as a period, all daily SM data with a spatial resolution of 1 km from 2015 to 2020 were predicted. In addition, to obtain a more complete time series of SM data, we used the model of the previous period when the number of valid samples was less than 100.

2.3.3.4 Evaluation method

The correlation coefficient (R) and the root mean square error (RMSE) were used to evaluate the accuracy of the regression model based on these machine learning methods (MLR, SVR, ANN, RF and XGB). They are calculated as:

$$R = \frac{Cov(SM_I, SM_P)}{\sqrt{Var(SM_I) Var(SM_P)}} \tag{5}$$

$$RMSE = \sqrt{\frac{1}{n}(SM_P - SM_I)^2} \tag{6}$$

where SM_I is the SMAP SM, SM_P is the corresponding SM predicted by the regression model, Cov represents the covariance function, Var is the variance, and n is the number of valid samples for SM_I or SM_P.

The regression model with the smallest average RMSE of training and test datasets was selected as the optimal model.

$$\overline{RMSE} = \frac{RMSE_{Training} + RMSE_{Test}}{2} \tag{7}$$

where $RMSE_{Training}$ and $RMSE_{Test}$ are the RMSE of the training set and test set for these models, respectively.

We used the selected optimal model with these surface variables with a resolution of 1 km within 16 days to simulate SM at 1 km resolution on the corresponding date. Taking 16 days as a period, we predicted all daily SM data with a spatial resolution of 1 km from 2015 to 2020. In addition, to obtain a more complete time series of SM data, we used the model of the previous period when the number of valid samples was less than 100.The in situ SM measurements were used to validate the downscaled results. In addition to R and RMSE, bias and unbiased RMSE (ubRMSE) were also used for accuracy evaluation. Bias indicates the overall level of overestimation or underestimation of simulation results. ubRMSE can eliminate the influence of deviation. They were calculated according to:

The in situ SM measurements were used to validate the downscaled results. In addition to R and RMSE, unbiased RMSE (ubRMSE) and bias were also calculated according to:

$$ubRMSE = \sqrt{\frac{1}{n}((SM_{In} - \overline{SM_{In}}) - (SM_D - \overline{SM_D}))^2} \tag{8\,10}$$

$$bias = \overline{SM_{In}} - \overline{SM_D} \tag{9\,11}$$

313 where SM_{In} is the in situ observed SM, SM_d is the downscaled SM of the corresponding grid, and n is the number of valid

314 samples for SM_{In} or SM_D.

315 **3. Results**

316 **3.1 Model comparison**

317 The daily SM from DOY 81 in 2015 to DOY 366 in 2020 were simulated producing 128 regression results every 16 days.

318 The correlation coefficient (R) and the root mean square error (RMSE) of each regression result for the training set and the

319 test set are shown in Fig. 4 and Fig. 5, respectively. According to Equation 9, among the 128 regression results, there were 114

320 from the XGB model, and 14 from RF.

321 For all models except MLR, R is greater than 0.6 and RMSE is less than 0.05 m³/m³ both for the training and the test set.

322 R greater than 0.6 and 0.8 indicates there is a reliable and strong correlation (Akoglu, 2018). It means that all methods except

323 MLR have reliable simulation accuracy. For the training set using XGB, Rs are all above 0.96, generally higher than for other

324 methods; Similarly, the RMSEs of XGB are all lower than 0.02 m³/m³, generally lower than those of other methods. The R of

325 RF is second only to that of XGB, and for several periods it is higher than for XGB; the RMSEs of RF are also generally lower

326 than 0.02 m³/m³ and are lower than those of XGB in several periods. SVR and ANN perform generally better in the cold

327 season, and worse in other seasons. In general, their results are inferior to those of XGB and RF. The simulation results of

328 MLR are relatively poor both in terms of RMSE and R.

329 ~~The daily SM from DOY 81 in 2015 to DOY 366 in 2020 were simulated producing 128 regression results every 16 days.~~

330 ~~According to Equation 3, among the 128 regression results, there were 123 from the XGB model, and 5 from RF (including~~

331 ~~DOY 2015241, 2016161, 2016209, 2017241 and 2017257).~~

332 ~~The correlation coefficient (R) and the root mean square error (RMSE) of each regression result for the training set and~~

333 ~~the test set are shown in Fig. 4 and Fig. 5, respectively. For all models, R is greater than 0.8 and RMSE is less than 0.1 both~~

334 ~~for the training and the test set. For the training set using XGB, Rs are all above 0.96, generally higher than for other methods;~~

335 ~~Similarly, the RMSEs of XGB are all lower than 0.02, generally lower than those of other methods. The R of RF is second~~

336 ~~only to that of XGB, and for some periods it is higher than for XGB; the RMSEs of RF are also generally lower than 0.02 and~~

337 ~~are lower than those of XGB in some periods. SVR and ANN perform better in the cold season, and worse in other seasons.~~

338 ~~In general, their results are inferior to those of XGB and RF. The simulation results of MLR are relatively poor both in terms~~

339 ~~of RMSE and R.~~

340 The results of the test set show that XGB, RF and SVR perform better than ANN and MLR~~, and are better in the cold~~

341 ~~season~~. Table 4 shows the average RMSE and R values of the training and test sets over all periods, and the performance order

342 of the model can be obtained as XGB>RF>SVR >ANN >MLR. In addition, there are seasonal variations in R and RMSE both

343 for training and test sets. Moreover, the evaluation accuracy was generally better in the cold season, when sample sizes were

344 smaller.

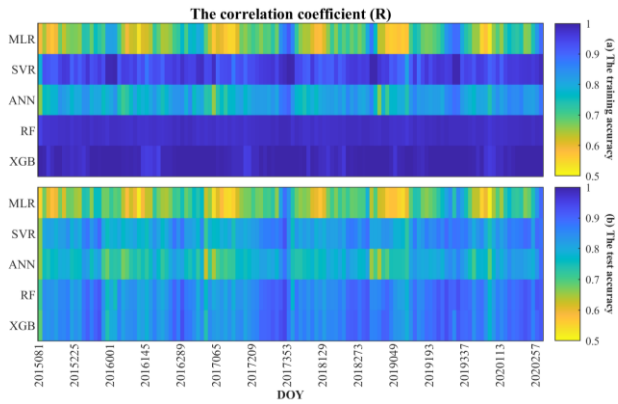



Figure 4: The correlation coefficient (R) of the models (MLR, SVR, ANN, RF and XGB) on different periods: (a) The training accuracy; (b) The test accuracy.

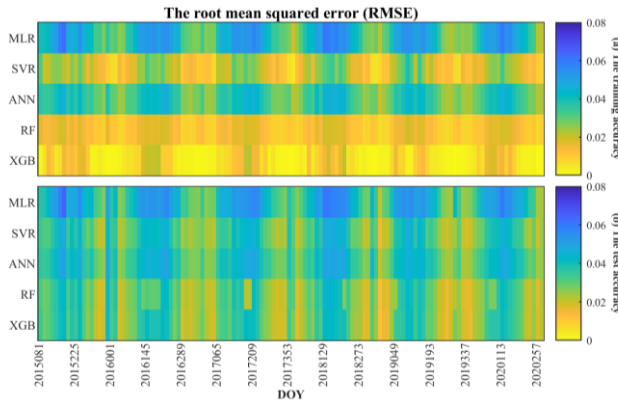



Figure 5: The root mean square error (RMSE) of the models (MLR, SVR, ANN, RF and XGB) for different periods: (a) The training accuracy; (b) The test accuracy.

Table 4: Accuracy of the models based on correlation coefficient (R) and root mean square error (RMSE)

	Model	MLR	SVR	ANN	RF	XGB	Combination
Training set	R	0.688	0.943	0.864	0.978	0.991	0.992
	RMSE (m^3/m^3)	0.042	0.019	0.028	0.013	0.007	0.007
Test set	R	0.675	0.824	0.660	0.857	0.861	0.861
	RMSE (m^3/m^3)	0.043	0.033	0.047	0.030	0.029	0.028

Model		MLR	SVR	ANN	RF	XGB
Training set	R	0.688	0.843	0.864	0.979	0.985
	RMSE	0.042	0.032	0.028	0.013	0.010
Test set	R	0.677	0.843	0.660	0.857	0.861
	RMSE	0.043	0.029	0.047	0.030	0.029

3.2 Comparison with the in situ data and precipitation

The downscaled 1 km gridded SM were compared with the in situ SM observations of the Maqu Network and Babao Network (Fig. 6). Due to the difference in sensors, soil depth and measurement scale (point observation in case of the in situ measured SM and 1 km grid for the downscaled SM), there is a certain deviation between in situ observation data and the downscaled gridded SM data. The downscaled SM of most sites at the Maqu Network and Babao Network are highly correlated with the in situ measured SM (R>0.6). In the Maqu Network, the ubRMSEs with an average of 0.057 m^3/m^3 are all less than 0.090 m^3/m^3, and the bias ranges from -0.10 to 0.22 m^3/m^3. In the Babao Network, the average ubRMSE of all sites is 0.081 m^3/m^3, and some of them exceed 0.1 m^3/m^3. In addition, their bias ranges from -0.07 to 0.45 m^3/m^3. It means that the validation accuracy of Babao Network is generally lower than Maqu Network. That may be mainly because the measured soil depth at the Babao Network is 4 cm, which means that there could be a systematic error between the datasets. Therefore, the validation accuracy should mainly refer to the evaluation accuracy of Maqu network.

To evaluate the performance of the downscaling approach, the downscaled 1 km gridded SM were compared with the in situ SM observations. The SM before and after downscaling were both compared with the in situ SM data of the Maqu Network and Babao Network (Fig. 6). Due to the difference in sensors, soil depth and measurement scale (point observation in case of the in situ measured SM and 1 km grid for the downscaled SM), there is a certain deviation between in situ observation data and the downscaled gridded SM data. The downscaled SM of most sites at the Maqu Network are highly correlated with the in situ measured SM (R>0.6). The ubRMSEs with an average of 0.049 m^3/m^3 are all less than 0.073 m^3/m^3, and the bias ranges from -0.10 to 0.15 m^3/m^3. The comparative results of the Babao Network are not as good as that of the Maqu Network. The SM data of most sites at the Babao Network have larger ubRMSE and bias, and the correlation coefficients between in situ observed SM and the downscaled SM are generally lower. That may be mainly because the measured soil depth at the Babao Network is 4 cm, which means that there could be a systematic error between the datasets.

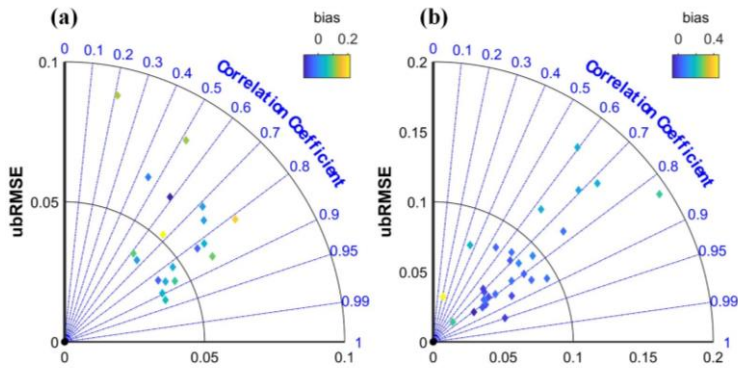
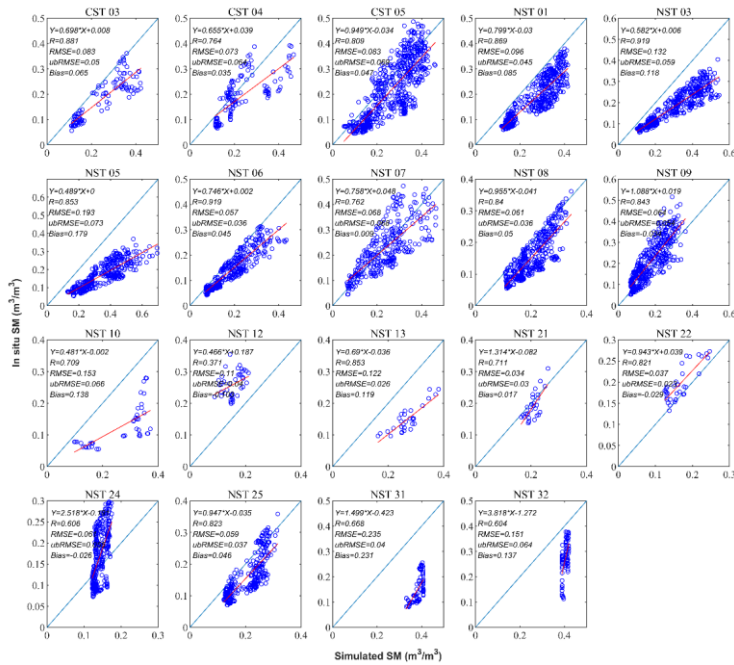



Figure 6: The relationships between in situ SM and downscaled SM. (a) Maqu Network; (b) Babao Network.

To better understand the reason for these poor results, the scatter plots comparing the two sets of data were drawn. Figure 7 shows the results of the 19 sites of the Maqu Network. All four statistical metrics, namely, R, RMSE, ubRMSE and bias were calculated, and their fitting line of the scatter was also plotted. Not surprisingly, the relationship is generally improved where there are more valid data. It means that the validation effect of in situ observations is affected by the amount of data. The same conclusion can be drawn through 29 sites at the Babao Network (Fig. S1).Not surprisingly, the relationship is generally improved where there are more points. The same conclusion can be drawn according to Fig. S1, which shows the comparative results of 29 sites at the Babao Network.

383

384 **Figure 7: Comparison between the downscaled SM and in situ SM of the Maqu Network.**

385   All SM products are compared with in situ SM. Figure 8 shows a significantly higher correlation between the downscaled

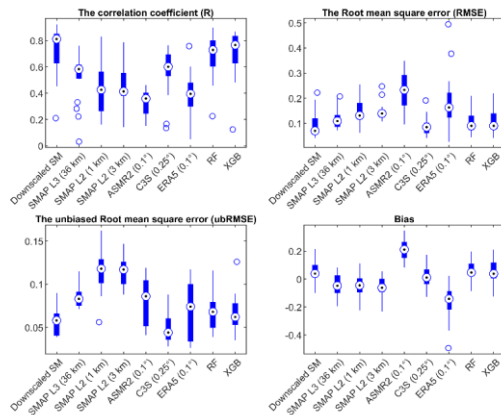
386 SM and in situ SM of the Maqu Network. The median ubRMSE of the downscaled SM is the smallest, and its RMSE is second

387 only to the C3S (0.25°) product. The bias of the downscaled SM is higher than that of some products, even higher than the

388 original SMAP L3 (36 km) data. Almost the same results can be obtained from in situ observations of Babao Network (Fig.

389 S2). The difference is that the bias of the downscaled SM is lower than the result of SMAP L3 (36 km). Compared with the

390 RF-based and the XGB-based downscaled SMs, the downscaled SM with multiple machine learning approaches performed

391 better, especially R and ubRMSE.

392

393

394

395   The observed SM of sites with a greater number of observed data were compared with these gridded SM data at different

396 resolutions and precipitation. ~~Figure 8~~Figure 9 shows the temporal variations of these SM at four sites 2016-2017. The

397 relationship between in situ observed SM and precipitation at all four sites is very consistent, showing annual fluctuation. The

398 greater SM corresponds to more precipitation during the hot season, and the smaller SM corresponds to less precipitation

399 during the cold season.

400   Except for GCOMW/ASMR2 SM, the variation trends of these acquired gridded SM and the downscaled SM are basically

401 the same despite the large difference in spatial resolution. GCOMW/ASMR2 significantly undestimates SM compared to other

402 products. Both the SMAP L2 SM at 1 km and 3 km may be overestimated (CST05) and may also be underestimated (WSN18)

403 compared with in situ observations.~~Both the SMAP L2 SM at 1 km and 3 km are overestimated compared with in situ~~

404 ~~observations.~~ Moreover, SMAP L2 SM has some valid data mainly on hot days and almost no valid data during cold seasons.

405 The peak values of the ERA5 SM are close to those of the in situ observations, but the low values are overestimated. The C3S

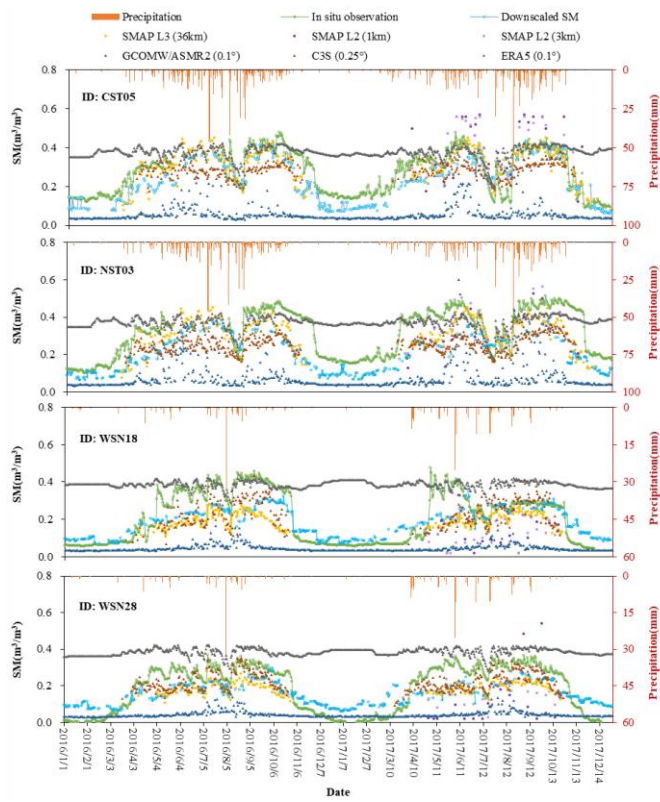
406 SM is similar to the 36 km SMAP SM, and its peak values are simulated more accurately, while the minimum values have

407 little valid data. Compared with the original data (36 km SMAP L3), the downscaled SM has a more complete time series,

408 especially during the cold season. The downscaled SM data almost all match well with the in situ measured SM data, and all

409 of them are also consistent with the precipitation. The difference between the downscaled SM and the in situ measured SM is

410 mainly reflected in the magnitude of the variation, which is probably due to the difference in spatial resolution.

411

412 **~~Figure 8~~Figure 9: Time series of the in situ observed SM, the downscaled SM, the acquired gridded SM products and daily**

413 **precipitation at the four selected SM sites (From Maqu Network and Babao Network, respectively) in 2016-2017~~.~~**

414 **3.3 Mapping of the downscaled SM**

415   SM varies greatly in different months in desertified areas. ~~Figure 9~~Figure 10 shows the average SM in each month in the

416 study area. The SM shows a monthly change pattern, and the values from June to September are bigger than in other months,

417 especially in southern Qinghai Province, eastern Inner Mongolia Province, and western Xinjiang Province~~, which is consistent~~

418 ~~with the process of vegetation growth.~~ The SM in some areas is low throughout the year, such as in the Tarim Basin of

419 Xinjiang Province, western Inner Mongolia Province and most of Gansu Province.

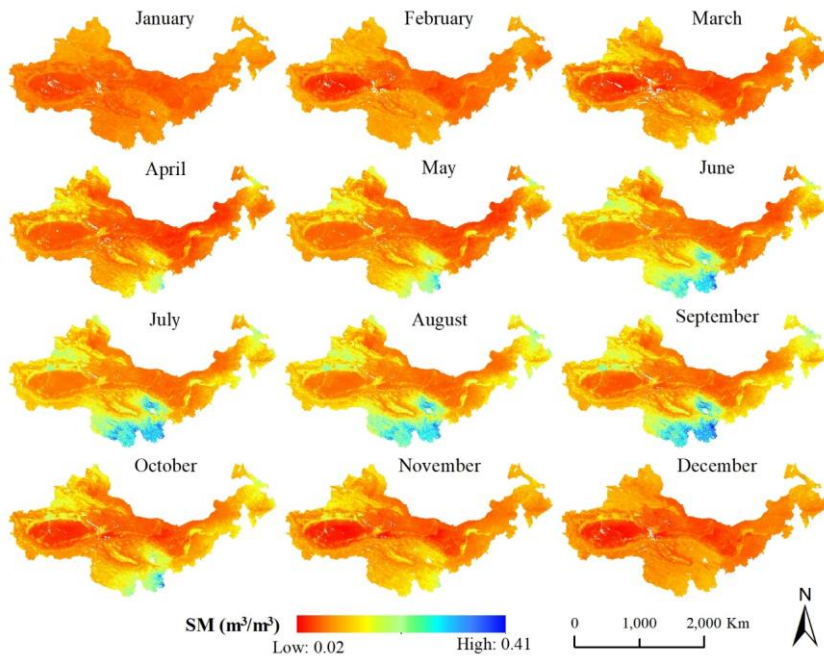
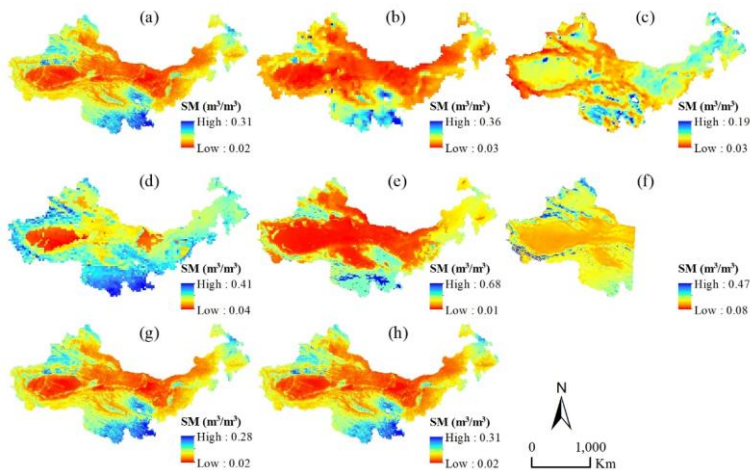



Figure 10: Monthly average SM in the study area.

The annual average SM was also calculated (Fig. ~~S2~~S3). Compared with the monthly average SM, the annual average SM changed significantly less.~~Overall, there is little variation in SM in different years.~~ Further, we compared the spatial patterns of the downscaled SM with the gridded SM products with different resolutions. ~~Figure 10~~Figure 11 shows the daily average SM of these products from 2015 to 2020. The spatial patterns of the downscaled SM and 36 km SMAP SM are basically consistent, but the downscaled data show better details in some areas such as near rivers. The overall values of GCOMW SM are relatively small, and exhibit some obvious errors in some areas. For example, SM in the Tarim Basin is higher than in the surrounding area, which is completely inconsistent with other SM data. The spatial pattern of the C3S SM is close to the downscaled SM and the 36 km SMAP SM, but some details are not presented. For example, SM in the Hetao Plain along the Yellow River is much higher than that in its surrounding area, which can be found in the downscaled SM and the SMAP SM, but not in the C3S SM. There are obvious errors in the results of ERA5. The average SM is significantly overestimated in the southern part of the study area, and underestimated in some areas in the northern of the study area.~~The average SM of the ERA5 products is polarised. In some areas the values are very large, and in some small areas they are very small.~~ The FLDAS SM has high resolution, and its overall spatial pattern is relatively consistent with the downscaled SM and

435   36 km SMAP SM. The difference is that the FLDAS SM is significantly larger in higher elevation areas of the west than in
436   other regions, which is quite different from other products. This suggests that the FLDAS SM may be overestimated in these
437   regions. In addition, FLDAS SM does not show wetter soil along the river. The spatial patterns of the RF-based and XGB-
438   based downscaled SMs are both close to that of the downscaled SM with multiple machine learning approaches, however, the
439   the maximum SM based on RF is smaller than the results based on XGB and multi-model combination.

440
441   ~~Figure 10~~Figure 11: Daily average SM from 2015-2020 in the study area. (a)-(f) are the downscaled SM (1 km), SMAP L3 SM (36
442   km), GCOMW/ASMR2 SM (0.1°), C3S SM (0.25°), ERA5 SM (0.1°) ~~and~~, FLDAS SM (0.1°) and the RF-based downscaled SM (1
443   km), respectively.

444   To better demonstrate the differences in SM, a case of the Mu Us Desert was selected (~~Fig. 11~~Fig. 12). The Mu Us Desert
445   is located in a semi-arid area with annual average precipitation of generally less than 400 mm, decreasing gradually from
446   southeast to northwest. The main types of land cover are grassland and sandy land, and the salinization is serious in a few
447   areas. Desertification has been severe for a long time in the past but ~~has been~~ significantly reversed with artificial afforestation
448   in recent years.
449   SM shows an overall trend of gradual decrease from the southeast to the northwest (~~Fig. 11~~Fig. 12 (b)~(g)), which is
450   consistent with the distribution of precipitation. The average SM of the same location changes little from year to year. Overall,
451   it is relatively large in 2018 and relatively small in 2015, which is also roughly consistent with annual precipitation patterns.
452   Land cover types also have a certain influence on the spatial difference of SM. The northwestern portion of the Mu Us Desert
453   is mainly grassland, which is strongly dependent on precipitation (Fig~~.~~. ~~11~~12 (h)). The southeastern area is mainly cultivated
454   land and is less affected by precipitation as it relies on pumping groundwater rather than natural precipitation (Fig~~.~~. ~~11~~12 (j)).

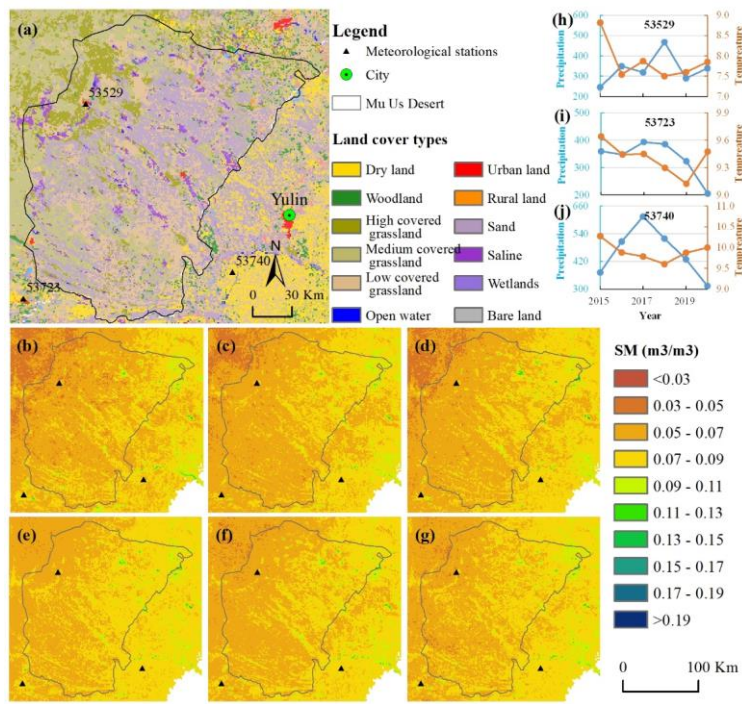



Figure 11Figure 12: Soil moisture estimated for the Mu Us Desert. (a) Land cover distribution over the study area; (b)-(g) annual average SM from 2015-2020; (h)-(j) annual precipitation and annual average temperature of three sites (53529, 53723 and 53740), whose surroundings are mainly grassland, cultivated land, and cultivated land, respectively.

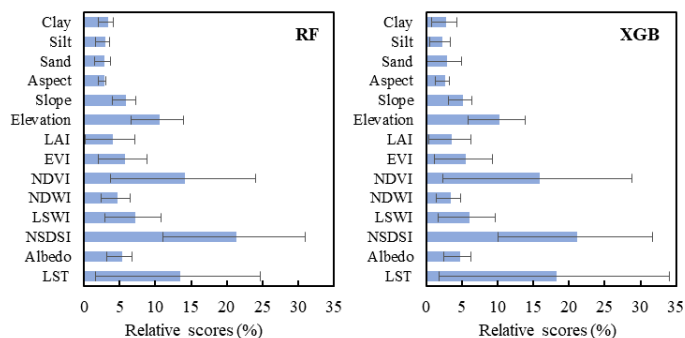
4. Discussion

4.1 Variable importance assessmentRegression variable importance

The selection of variables is an important step of a nonlinear regression model. The importance analysis of the variables carried out for this research found that a larger number of variables can improve the regression effect of these models to some extent. Due to the variables obtained in this study come from multiple data sources, their preprocessing may affect the construction of regression models and their relationship with SM. Moreover, variables collinearity and hyperparameters also affects the importance relationship of variables. This average result of the 128 regression models can give a relatively

466   reasonable result. ~~Figure 12~~Figure 13 shows the average importance scores of each variable for the RF and XGB models across
467   all available days. The importance scores of different variables in the RF based model and the XGB based model are similar.
468   LST and surface albedo both affect surface energy exchange and partition. LST is a ~~very~~n important variable in both models,
469   which is consistent with the study of Zhao et al. (2018). NSDSI is the most sensitive soil moisture index compared to LSWI
470   and NDWI, which was demonstrated in Yue et al. (2019). Topographical factors also exhibit importance on SM, especially
471   elevation. NDVI is more sensitive to vagetation index than EVI and LAI. However, their effect was smaller than that of soil
472   moisture index. It indicates that the SM inversion method based only on LST and VI is inadequate. The influence of soil texture
473   (sand, silt and clay) is relatively weak~~, but it cannot be completely ignored~~.

474   The standard deviation of the importance scores of each variable is shown with error bars in Fig. 13~~2~~. Its changes are
475   mainly affected by the samples used in the regression model and the temporal variations in surface variables. For static
476   variables such as soil structure and topographic factors, the changes in their importance scores mainly depend on the number
477   and the location of the samples. ~~Figure 12~~Figure 13 also shows that their standard deviation is relatively small. Compared with
478   static variables, the standard deviation of the importance scores of dynamic variables is significantly larger, especially for LST
479   and LAI. This indicates that it is not reliable to construct a single regression model for a long time series.

480   In general, the variable importance analysis suggests that the selected variables are suitable for the construction of the
481   regression model. Moreover, choosing 16 days as a time period to build a regression model benefits from obtaining a sufficient
482   number of samples, especially since the surface variables were found still unchanged during these intervals.

483

484   **~~Figure 12~~Figure 13: The average importance scores of variables for the RF based approach and XGB based approach. Note: The**
485   **importance scores are presented by increase in Node Purity (IncNodePurity) ~~IncNodePurity~~ where the sum value is normalized for**
486   **the RF model; The XGB model uses Gain to reflect the weight of variables.**

487   **4.2 Advantages of model combination**

Both RF and ANN have been applied to downscale remote sensed SM so far, especially RF (Zhao et al., 2018; Qu et al., 2019; Hu et al., 2020). This study showed that the simulation results of ANN have greater uncertainty, and the accuracy is generally worse than that of RF (Figs. 4 and 5). The RF algorithm shows a good simulation ability, but in comparison, the XGB algorithm also has a corresponding effect or even higher. We also compared our simulation results combining multiple models and the RF-based simulation results. The results showed that the combined products have higher accuracy than the RF-based products, which is mainly reflected in the relatively more reasonable simulation of peaks and valleys (Table 4 and Fig 11). MLR has the worst effect compared to the other four models, which is likely to be affected by variable collinearity. In fact, many algorithms, especially linear ones, exhibit more or less poor robustness when there is high collinearity between variables (Dormann et al., 2013; Cammarota and Pinto, 2021). However, fewer explanatory variables also often means less ability to explain target variables. Several studies have shown that ensemble tree algorithms such as RF and XGB are generally not affected by variable collinearity (Tomaschek et al., 2018; Chen et al., 2020; Feng et al., 2021).

The simulation results of long time series will inevitably suffer the interference of various noises. A combination of multiple methods can reduce overfitting and uncertainties for the simulation of long time series (Zanotti et al., 2019; Yu et al., 2021). The five methods (MLR, SVR, ANN, RF and XGB) in this study have indicated the potential flaws of a single model. Although the XGB model XGB generally perform better than other models, it still has still some shortcomings. As it can be seen from Figs. 4 and 5, compared with the training accuracy, the test accuracy of the XGB model is significantly reduced in several periods. This means that the simulation results of the XGB model is likely to have a certain degree of overfitting. In contrast, the difference between training accuracy and test accuracy of the RF model is even smaller. It showed better stability than XGB at several some periods (Figs. 4 and 5). The training accuracy of MLR and SVR has a small difference from the test accuracy, but the overall accuracy is obviously lower, which is not suitable for remote sensing SM prediction (Table 4), which might be due to variable collinearity. Some studies have also proved that SVR may also perform better than some ensemble algorithms (Yu et al., 2012; Fan et al., 2018). The fitting effect of ANN varies greatly in different periods, indicating that its generalization is lower than other models (Piotrowski and Napiorkowski, 2013). In general, the XGB and RF models provide the best combination of prediction accuracy and stability.

4.3 Analysis of the relationship with precipitation and temperature

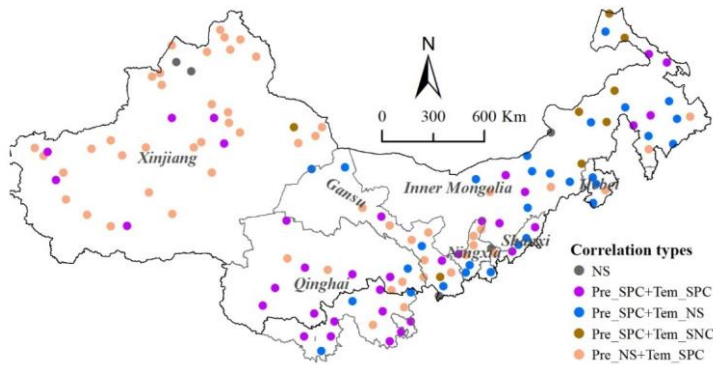
Unlike predictors such as LST and NDVI that reflect SM status, climatic factors are are key drivers of SM variability. To evaluate the impact of precipitation and temperature on SM, we performed a partial correlation analysis on the data of all meteorological stations. Precipitation and temperature are important factors affecting SM. To evaluate the impact of precipitation and temperature on SM, we performed a partial correlation analysis on the data of all meteorological stations. Figure 13 Figure 14 shows that SM is mainly positively correlated with precipitation and temperature, and a few regions are significantly negatively correlated with temperature. In terms of spatial distribution, SM of the sites in the eastern region (including Inner Mongolia Province, Hebei Province and Shanxi Province) is mainly significantly affected by precipitation.

520 Due to the influence of glaciers and snowmelt, the SM of the sites in the western region (Xinjiang Province and Gansu Province)

521 is more affected by temperature. In addition, the number of sites with significant positive correlation with precipitation and

522 temperature is the largest in Qinghai Province. This indicates that precipitation and temperature in the eastern part of the

523 Tibetan Plateau both have a great influence on SM.

524

525 ~~Figure 13~~Figure 14: Partial correlation between monthly downscaled SM and precipitation and temperature (Pre: precipitation;

526 Tem: temperature; NS: Not significance; SPC: Significantly positive correlation; SNC: Significantly negative correlation).

4.4 Uncertainty and Prospects

528 While this study greatly improved the spatial resolution of SM data from 2015-2020 in the desertification areas of North

529 China by downscaling SMAP SM products, it still presents some shortcomings. Due to the influence of snow, ice and frozen

530 ground, the number of valid SMAP pixels during cold seasons is very small, that makes the number of available samples is

531 limited. With a period of 16-day, the number of valid samples may still be less than 100 during cold seasons (Fig. 3). The

532 sample size affects the simulation accuracy. Figures 4 and 5 show that there are seasonal variations in R and RMSE, which is

533 likely to be affected by the sample size (Figure 3). In general, a larger sample size often mean more efficient sampling and

534 more reliable results, but not necessarily better evaluation accuracy. Likewise, insufficient samples can sometimes have good

535 evaluation accuracy, although the results are less reliable. In order to reduce the error caused by insufficient samples, this study

536 replaced the periods with less than 100 samples with the model of the previous periods. For this reason, the simulation results

537 sometimes perform poorly during cold seasons (Fig. 9). In addition, the upscaling (from 1 km to 36 km resolution) of surface

538 variables also has a certain impact on the accuracy of the model.

539 Our products have a good correlation with the in situ observation data. Although there are deviations, the results are

540 relatively reasonable. However, in situ observed SM data are limited in their representation of the entire 1 km ×1 km grid,

541 which adds to the uncertainty of our product. Figure 6 shows that the evaluation accuracy of different points varies greatly.

542 Through the investigation of past literature and our study, it is found that the relationship between in situ observation data and

remote sensing SM has great uncertainty due to the influence of scale (Zeng et al., 2015; Abbaszadeh et al., 2019; Bai et al., 2019; Liu et al., 2019; Zhang et al., 2020). In addition, due to instrument accuracy and climate change, there are also large errors in the in situ observation data, especially at low temperatures. The in situ observed SM data obtained in this paper are relatively limited, and their spatial distribution is concentrated in a certain part of of the study area, which is not representative. It increases the uncertainty of the simulation results. In order to verify the accuracy of the data as much as possible, this study also selected several sets of gridded SM products for comparison. The results showed that our products perform better in temporal variability and spatial patterns (Figs. 9 and 11).

While this study greatly improved the spatial resolution of SM data from 2015-2020 in the desertifying areas of North China by downscaling SMAP SM products, it still presents some shortcomings. For example, due to the image quality and coverage of SMAP and the impact of noise from clouds on the MODIS products, the number of valid samples for a 16-day period may still be less than 100 points. This study replaced the periods with less than 100 samples with the model of the previous periods. Due to the limited number of available samples, the simulation in the cold season is relatively poor (Fig. 8). In addition, the upscaling (from 1 km to 36 km resolution) of surface variables also has a certain impact on the accuracy of the model. The in situ observed SM data obtained in this paper are relatively limited, and their spatial distribution is concentrated in a certain part of of the study area, which is not representative. It increases the uncertainty of the simulation results. In order to verify the accuracy of the data as much as possible, we compared this product with the existing more reliable gridded SM products, and the results showed that our product showed certain advantages in both time series and spatial distribution (Figs .8 and 10).(Zeng et al., 2015; Abbaszadeh et al., 2019; Bai et al., 2019; Liu et al., 2019; Zhang et al., 2020)(Zhao et al., 2018; Qu et al., 2019; Hu et al., 2020)

The Chinese government focuses on desertification reduction through afforestation and the establishment of grasslands. SM data with high temporal and spatial resolution can provide a reference for the next steps of revegetation.

5 DCode and data availability

The codes mainly used in this paper mainly includes sample selection, the building of the optimal regression model and the result prediction. The downscaled daily SM dataset at 1 km spatial resolution is available at https://doi.org/10.6084/M9.FIGSHARE.16430478.V5 (Rao et al., 2021). The data maps are all provided in Geotiff format, and the value has expanded 10, 000 times to make them easier to store. The filenames reflect the production date in Julian Day format.

The codes mainly used in this paper mainly includes sample selection, the building of the optimal regression model and the result prediction. These codes based on the R language can be found in the supplementary documents. The downscaled daily SM dataset at 1 km spatial resolution is available at https://doi.org/10.6084/M9.FIGSHARE.16430478.V5 (Rao et al., 2021). The data maps are all provided in Geotiff format, and the value has expanded 10, 000 times to make them easier to store. The filenames reflect the production date in Julian Day format.

6 Conclusions

In this study, an approach~~a framework~~ was proposed for downscaling 36 km SMAP SM products using MODIS optical products and other surface variables (mainly topographic data and soil data) based on multiple machine learning methods. Overall, the regression performance of the five methods is, in order: XGB>RF>SVR>ANN>MLR. Compared with MLR, SVR and ANN, XGB and RF have much better ~~regression~~ accuracy, and they were used in combination to produce daily 1 km downscaled SM in a period of 16 days. The validation shows that the downscaled SM are highly related to most in situ measured SM. The ubRMSE with an average of 0.~~049~~057 m^3/m^3 is generally less than 0.09~~0~~073 m^3/m^3 at the Maqu Network. Time series of SM data from in situ observation sites are also compared. The results show that the downscaled SMs are highly related to SMAP SMs, and provide a more complete time series and match better with the in situ measured SM. Compared with some commonly used gridded SM products such as SMAP L2 (l km or 3 km), GCOMW/ASMR2, C3S, ERA5 and FLDAS SMs, the downscaled SM data not only have higher spatial resolution, but also have a more reliable accuracy whether in time series or spatial distribution.

The maps of downscaled SM show larger values from June to September, which coincides with the vegetation growing season. The difference in annual mean SM is small. Spatially, SM is relatively large in Qinghai Province and in northeastern Inner Mongolia, especially in summer. In arid areas such as the Tarim Basin, SM is relatively small throughout the year. Moreover, precipitation and temperature both have a great influence on SM in the study area. Precipitation has a greater impact on SM in the eastern part of the study area, while the effect of temperature appears to be more pronounced in the west.

This approach makes it possible to more accurately assess the soil moisture status in the study area. The results can support regional agricultural planting and revegetation efforts and can be applied to limit desertification in other areas in the future.

Author contributions. FW and PR designed the research, developed the methodology, performed the analysis, and wrote the paper; YW, YL, XW, and ZW edited and revised the paper.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This work was supported by the National Key Research and Development Program of China (2018YFC0408103), the National Pilot Project for Ecological Protection and Restoration of Mountains, Rivers, Forests, Farmlands, Lakes and Grasslands (Grant No. WR0203A552018), and the Desertification Monitoring Project of National Forestry and Grass Administration (Grant No. 2020062012). We thank all data providers and the anonymous reviewers for their detailed and constructive comments.

References

606 Abbaszadeh, P., Moradkhani, H., and Zhan, X.: Downscaling SMAP Radiometer Soil Moisture Over the CONUS Using
607 an Ensemble Learning Method, Water Resour. Res., 55, 324–344, https://doi.org/10.1029/2018WR023354, 2019.

608 Achieng, K. O.: Modelling of soil moisture retention curve using machine learning techniques: Artificial and deep neural
609 networks vs support vector regression models, Computers & Geosciences, 133, 104320,
610 https://doi.org/10.1016/j.cageo.2019.104320, 2019.

611 Ågren, A. M., Larson, J., Paul, S. S., Laudon, H., and Lidberg, W.: Use of multiple LIDAR-derived digital terrain indices
612 and machine learning for high-resolution national-scale soil moisture mapping of the Swedish forest landscape, Geoderma,
613 404, 115280, https://doi.org/10.1016/j.geoderma.2021.115280, 2021.

614 Bai, J., Cui, Q., Zhang, W., and Meng, L.: An Approach for Downscaling SMAP Soil Moisture by Combining Sentinel-
615 1 SAR and MODIS Data, Remote Sensing, 11, 2736, https://doi.org/10.3390/rs11232736, 2019.

616 Belgiu, M. and Drăguț, L.: Random forest in remote sensing: A review of applications and future directions, 114, 24–31,
617 https://doi.org/10.1016/j.isprsjprs.2016.01.011, 2016.

618 Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD
619 International Conference on Knowledge Discovery and Data Mining, KDD '16: The 22nd ACM SIGKDD International
620 Conference on Knowledge Discovery and Data Mining, San Francisco California USA, 785–794,
621 https://doi.org/10.1145/2939672.2939785, 2016.

622 Chen, Y., Feng, X., and Fu, B.: An improved global remote-sensing-based surface soil moisture (RSSSM) dataset
623 covering 2003–2018, Earth Syst. Sci. Data, 13, 1–31, https://doi.org/10.5194/essd-13-1-2021, 2021.

624 De Santis, D., Biondi, D., Crow, W. T., Camici, S., Modanesi, S., Brocca, L., and Massari, C.: Assimilation of Satellite
625 Soil Moisture Products for River Flow Prediction: An Extensive Experiment in Over 700 Catchments Throughout Europe,
626 Water Res, 57, https://doi.org/10.1029/2021WR029643, 2021.

627 Del Frate, F., Ferrazzoli, P., and Schiavon, G.: Retrieving soil moisture and agricultural variables by microwave
628 radiometry using neural networks, Remote Sensing of Environment, 84, 174–183, https://doi.org/10.1016/S0034-
629 4257(02)00105-0, 2003.

630 Demarchi, L., Kania, A., Ciężkowski, W., Piórkowski, H., Oświecimska-Piasko, Z., and Chormański, J.: Recursive
631 Feature Elimination and Random Forest Classification of Natura 2000 Grasslands in Lowland River Valleys of Poland Based
632 on Airborne Hyperspectral and LiDAR Data Fusion, Remote Sensing, 12, 1842, https://doi.org/10.3390/rs12111842, 2020.

633 Elshorbagy, A. and Parasuraman, K.: On the relevance of using artificial neural networks for estimating soil moisture
634 content, Journal of Hydrology, 362, 1–18, https://doi.org/10.1016/j.jhydrol.2008.08.012, 2008.

635 Fan, J., Yue, W., Wu, L., Zhang, F., Cai, H., Wang, X., Lu, X., and Xiang, Y.: Evaluation of SVM, ELM and four tree-
636 based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates
637 of China, Agricultural and Forest Meteorology, 263, 225–241, https://doi.org/10.1016/j.agrformet.2018.08.019, 2018.

638 Fan, J., Zheng, J., Wu, L., and Zhang, F.: Estimation of daily maize transpiration using support vector machines, extreme
639 gradient boosting, artificial and deep neural networks models, Agricultural Water Management, 245, 106547,
640 https://doi.org/10.1016/j.agwat.2020.106547, 2021.

641   Fang, B. and Lakshmi, V.: Soil moisture at watershed scale: Remote sensing techniques, Journal of Hydrology, 516, 258–
642   272, https://doi.org/10.1016/j.jhydrol.2013.12.008, 2014.

643   Fang, B., Lakshmi, V., Bindlish, R., Jackson, T. J., Cosh, M., and Basara, J.: Passive Microwave Soil Moisture
644   Downscaling Using Vegetation Index and Skin Surface Temperature, Vadose Zone Journal, 12, vzj2013.05.0089er,
645   https://doi.org/10.2136/vzj2013.05.0089er, 2013.

646   Gu, Y., Hunt, E., Wardlow, B., Basara, J. B., Brown, J. F., and Verdin, J. P.: Evaluation of MODIS NDVI and NDWI for
647   vegetation drought monitoring using Oklahoma Mesonet soil moisture data, Geophys. Res. Lett., 35, L22401,
648   https://doi.org/10.1029/2008GL035772, 2008.

649   Hu, F., Wei, Z., Zhang, W., Dorjee, D., and Meng, L.: A spatial downscaling method for SMAP soil moisture through
650   visible and shortwave-infrared remote sensing data, Journal of Hydrology, 590, 125360,
651   https://doi.org/10.1016/j.jhydrol.2020.125360, 2020.

652   Im, J., Park, S., Rhee, J., Baik, J., and Choi, M.: Downscaling of AMSR-E soil moisture with MODIS products using
653   machine learning approaches, Environ Earth Sci, 75, 1120, https://doi.org/10.1007/s12665-016-5917-6, 2016.

654   Kang, J., Jin, R., Li, X., Ma, C., Qin, J., and Zhang, Y.: High spatio-temporal resolution mapping of soil moisture by
655   integrating wireless sensor network observations and MODIS apparent thermal inertia in the Babao River Basin, China,
656   Remote Sensing of Environment, 191, 232–245, https://doi.org/10.1016/j.rse.2017.01.027, 2017.

657   Lievens, H., Verhoest, N. E. C., De Keyser, E., Vernieuwe, H., Matgen, P., Álvarez-Mozos, J., and De Baets, B.: Effective
658   roughness modelling as a tool for soil moisture retrieval from C- and L-band SAR, Hydrol. Earth Syst. Sci., 15, 151–162,
659   https://doi.org/10.5194/hess-15-151-2011, 2011.

660   Liu, J., Chai, L., Lu, Z., Liu, S., Qu, Y., Geng, D., Song, Y., Guan, Y., Guo, Z., Wang, J., and Zhu, Z.: Evaluation of
661   SMAP, SMOS-IC, FY3B, JAXA, and LPRM Soil Moisture Products over the Qinghai-Tibet Plateau and Its Surrounding Areas,
662   Remote Sensing, 11, 792, https://doi.org/10.3390/rs11070792, 2019.

663   Liu, J., Chai, L., Dong, J., Zheng, D., Wigneron, J.-P., Liu, S., Zhou, J., Xu, T., Yang, S., Song, Y., Qu, Y., and Lu, Z.:
664   Uncertainty analysis of eleven multisource soil moisture products in the third pole environment based on the three-corned hat
665   method, Remote Sensing of Environment, 255, 112225, https://doi.org/10.1016/j.rse.2020.112225, 2021.

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

669   Ma, M., Zhao, G., He, B., Li, Q., Dong, H., Wang, S., and Wang, Z.: XGBoost-based method for flash flood risk
670   assessment, Journal of Hydrology, 598, 126382, https://doi.org/10.1016/j.jhydrol.2021.126382, 2021.

671   Mallick, K., Bhattacharya, B. K., and Patel, N. K.: Estimating volumetric surface moisture content for cropped soils using
672   a soil wetness index based on surface temperature and NDVI, Agricultural and Forest Meteorology, 149, 1327–1342,
673   https://doi.org/10.1016/j.agrformet.2009.03.004, 2009.

674   Meng, X., Mao, K., Meng, F., Shi, J., Zeng, J., Shen, X., Cui, Y., Jiang, L., and Guo, Z.: A fine-resolution soil moisture
675   dataset for China in 2002–2018, Geosciences – Geophysics, https://doi.org/10.5194/essd-2020-292, 2020.

676   Mohana, R. M., Reddy, C. K. K., Anisha, P. R., and Murthy, B. V. R.: Random forest algorithms for the classification of
677   tree-based ensemble, Materials Today: Proceedings, S2214785321008853, https://doi.org/10.1016/j.matpr.2021.01.788, 2021.

678   O'Neill, P., Entekhabi, D., Njoku, E., and Kellogg, K.: The NASA Soil Moisture Active Passive (SMAP) mission:
679   Overview, in: 2010 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2010 - 2010 IEEE International
680   Geoscience   and   Remote   Sensing   Symposium,   Honolulu,   HI,   USA,   3236–3239,
681   https://doi.org/10.1109/IGARSS.2010.5652291, 2010.

682   Peng, J., Loew, A., Merlin, O., and Verhoest, N. E. C.: A review of spatial downscaling of satellite remotely sensed soil
683   moisture: Downscale Satellite-Based Soil Moisture, Rev. Geophys., 55, 341–366, https://doi.org/10.1002/2016RG000543,
684   2017.

685   Peng, J., Albergel, C., Balenzano, A., Brocca, L., Cartus, O., Cosh, M. H., Crow, W. T., Dabrowska-Zielinska, K., Dadson,
686   S., Davidson, M. W. J., de Rosnay, P., Dorigo, W., Gruber, A., Hagemann, S., Hirschi, M., Kerr, Y. H., Lovergine, F., Mahecha,
687   M. D., Marzahn, P., Mattia, F., Musial, J. P., Preuschmann, S., Reichle, R. H., Satalino, G., Silgram, M., van Bodegom, P. M.,
688   Verhoest, N. E. C., Wagner, W., Walker, J. P., Wegmüller, U., and Loew, A.: A roadmap for high-resolution satellite soil
689   moisture applications – confronting product characteristics with user requirements, 15, 2021.

690   Piles, M., Petropoulos, G. P., Sánchez, N., González-Zamora, Á., and Ireland, G.: Towards improved spatio-temporal
691   resolution soil moisture retrievals from the synergy of SMOS and MSG SEVIRI spaceborne observations, Remote Sensing of
692   Environment, 180, 403–417, https://doi.org/10.1016/j.rse.2016.02.048, 2016.

693   Piotrowski, A. P. and Napiorkowski, J. J.: A comparison of methods to avoid overfitting in neural networks training in
694   the case of catchment runoff modelling, Journal of Hydrology, 476, 97–111, https://doi.org/10.1016/j.jhydrol.2012.10.019,
695   2013.

696   Qu, Y., Zhu, Z., Chai, L., Liu, S., Montzka, C., Liu, J., Yang, X., Lu, Z., Jin, R., Li, X., Guo, Z., and Zheng, J.: Rebuilding
697   a Microwave Soil Moisture Product Using Random Forest Adopting AMSR-E/AMSR2 Brightness Temperature and SMAP
698   over the Qinghai–Tibet Plateau, China, Remote Sensing, 11, 683, https://doi.org/10.3390/rs11060683, 2019.

699   Rahimzadeh-Bajgiran, P., Berg, A. A., Champagne, C., and Omasa, K.: Estimation of soil moisture using optical/thermal
700   infrared remote sensing in the Canadian Prairies, ISPRS Journal of Photogrammetry and Remote Sensing, 83, 94–103,
701   https://doi.org/10.1016/j.isprsjprs.2013.06.004, 2013.

702   Rao, P., Jiang, W., Hou, Y., Chen, Z., and Jia, K.: Dynamic Change Analysis of Surface Water in the Yangtze River
703   Basin Based on MODIS Products, Remote Sensing, 10, 1025, https://doi.org/10.3390/rs10071025, 2018.

704   Rao, P., Wang, Y., Wang, F., Liu, Y., Wang, X., and Wang, Z.: Daily soil moisture mapping at 1-km resolution based on
705   SMAP data for areas affected by desertification in Northern China, https://doi.org/10.6084/M9.FIGSHARE.16430478.V3,
706   2021.

707   Sandholt, I., Rasmussen, K., and Andersen, J.: A simple interpretation of the surface temperature/vegetation index space
708   for assessment of surface moisture status, Remote Sensing of Environment, 79, 213–224, https://doi.org/10.1016/S0034-
709   4257(01)00274-7, 2002.

710   Shangguan, W., Dai, Y., Liu, B., Ye, A., and Yuan, H.: A soil particle-size distribution dataset for regional land and
711   climate modelling in China, Geoderma, 171–172, 85–91, https://doi.org/10.1016/j.geoderma.2011.01.013, 2012.

712   Shi, R., Xu, X., Li, J., and Li, Y.: Prediction and analysis of train arrival delay based on XGBoost and Bayesian
713   optimization, Applied Soft Computing, 109, 107538, https://doi.org/10.1016/j.asoc.2021.107538, 2021.

714   Sun, L., Sun, R., Li, X., Liang, S., and Zhang, R.: Monitoring surface soil moisture status based on remotely sensed
715   surface temperature and vegetation index information, Agricultural and Forest Meteorology, 166–167, 175–187,
716   https://doi.org/10.1016/j.agrformet.2012.07.015, 2012.

717   Wagner, W., Hahn, S., Kidd, R., Melzer, T., Bartalis, Z., Hasenauer, S., Figa-Saldaña, J., de Rosnay, P., Jann, A.,
718   Schneider, S., Komma, J., Kubu, G., Brugger, K., Aubrecht, C., Züger, J., Gangkofner, U., Kienberger, S., Brocca, L., Wang,
719   Y., Blöschl, G., Eitzinger, J., and Steinnocher, K.: The ASCAT Soil Moisture Product: A Review of its Specifications,
720   Validation Results, and Emerging Applications, metz, 22, 5–33, https://doi.org/10.1127/0941-2948/2013/0399, 2013.

721   Wang, G., Zhang, X., Yinglan, A., Duan, L., Xue, B., and Liu, T.: A spatio-temporal cross comparison framework for the
722   accuracies of remotely sensed soil moisture products in a climate-sensitive grassland region, Journal of Hydrology, 597,
723   126089, https://doi.org/10.1016/j.jhydrol.2021.126089, 2021.

724   Wang, S., Liu, S., Zhang, J., Che, X., Yuan, Y., Wang, Z., and Kong, D.: A new method of diesel fuel brands identification:
725   SMOTE   oversampling   combined   with   XGBoost   ensemble   learning,   Fuel,   282,   118848,
726   https://doi.org/10.1016/j.fuel.2020.118848, 2020.

727   Wang, T., Yang, D., Fang, B., Yang, W., Qin, Y., and Wang, Y.: Data-driven mapping of the spatial distribution and
728   potential changes of frozen ground over the Tibetan Plateau, Science of The Total Environment, 649, 515–525,
729   https://doi.org/10.1016/j.scitotenv.2018.08.369, 2019.

730   Wang, X., Xie, H., Guan, H., and Zhou, X.: Different responses of MODIS-derived NDVI to root-zone soil moisture in
731   semi-arid and humid regions, Journal of Hydrology, 340, 12–24, https://doi.org/10.1016/j.jhydrol.2007.03.022, 2007.

732   Yao, P., Shi, J., Zhao, T., Lu, H., and Al-Yaari, A.: Rebuilding Long Time Series Global Soil Moisture Products Using
733   the Neural Network Adopting the Microwave Vegetation Index, Remote Sensing, 9, 35, https://doi.org/10.3390/rs9010035,
734   2017.

735   Yu, H., Wu, Y., Niu, L., Chai, Y., Feng, Q., Wang, W., and Liang, T.: A method to avoid spatial overfitting in estimation
736   of   grassland   above-ground   biomass   on   the   Tibetan   Plateau,   Ecological   Indicators,   125,   107450,
737   https://doi.org/10.1016/j.ecolind.2021.107450, 2021.

738   Yu, Z., Liu, D., Lü, H., Fu, X., Xiang, L., and Zhu, Y.: A multi-layer soil moisture data assimilation using support vector
739   machines and ensemble particle filter, Journal of Hydrology, 475, 53–64, https://doi.org/10.1016/j.jhydrol.2012.08.034, 2012.

740   Yue, J., Tian, J., Tian, Q., Xu, K., and Xu, N.: Development of soil moisture indices from differences in water absorption
741   between shortwave-infrared bands, ISPRS Journal of Photogrammetry and Remote Sensing, 154, 216–230,
742   https://doi.org/10.1016/j.isprsjprs.2019.06.012, 2019.

743   Zanotti, C., Rotiroti, M., Sterlacchini, S., Cappellini, G., Fumagalli, L., Stefania, G. A., Nannucci, M. S., Leoni, B., and
744   Bonomi, T.: Choosing between linear and nonlinear models and avoiding overfitting for short and long term groundwater level
745   forecasting in a linear system, Journal of Hydrology, 578, 124015, https://doi.org/10.1016/j.jhydrol.2019.124015, 2019.

746   Zawadzki, J. and Kędzior, M.: Soil moisture variability over Odra watershed: Comparison between SMOS and GLDAS
747 data, International Journal of Applied Earth Observation and Geoinformation, 45, 110–124,
748 https://doi.org/10.1016/j.jag.2015.03.005, 2016.

749   Zeng, J., Li, Z., Chen, Q., Bi, H., Qiu, J., and Zou, P.: Evaluation of remotely sensed and reanalysis soil moisture products
750 over the Tibetan Plateau using in-situ observations, Remote Sensing of Environment, 163, 91–110,
751 https://doi.org/10.1016/j.rse.2015.03.008, 2015.

752   Zhang, P., Zheng, D., van der Velde, R., Wen, J., Zeng, Y., Wang, X., Wang, Z., Chen, J., and Su, Z.: Status of the Tibetan
753 Plateau observatory (Tibet-Obs) and a 10-year (2009–2019) surface soil moisture dataset, Hydrology and Soil Science –
754 Hydrology, https://doi.org/10.5194/essd-2020-209, 2020.

755   Zhao, W. and Li, A.: A Downscaling Method for Improving the Spatial Resolution of AMSR-E Derived Soil Moisture
756 Product Based on MSG-SEVIRI Data, Remote Sensing, 5, 6790–6811, https://doi.org/10.3390/rs5126790, 2013.

757   Zhao, W., Li, A., and Zhao, T.: Potential of Estimating Surface Soil Moisture With the Triangle-Based Empirical
758 Relationship Model, IEEE Trans. Geosci. Remote Sensing, 55, 6494–6504, https://doi.org/10.1109/TGRS.2017.2728815,
759 2017.

760   Zhao, W., Sánchez, N., Lu, H., and Li, A.: A spatial downscaling approach for the SMAP passive surface soil moisture
761 product using random forest regression, Journal of Hydrology, 563, 1009–1024, https://doi.org/10.1016/j.jhydrol.2018.06.081,
762 2018.

763