

Article

# Modeling Soil Moisture from Multisource Data by Stepwise Multilinear Regression: An Application to the Chinese Loess Plateau

Lina Yuan <sup>1</sup>, Long Li <sup>1,2</sup>, Ting Zhang <sup>1</sup>, Longqian Chen <sup>1,\*</sup>, Weiqiang Liu <sup>3</sup>, Sai Hu <sup>4</sup> and Longhua Yang <sup>5</sup>

<sup>1</sup> School of Public Policy and Management, China University of Mining and Technology, Daxue Road 1, Xuzhou 221116, China; lnyuan@cumt.edu.cn (L.Y.); long.li@cumt.edu.cn (L.L.); tingzhang@cumt.edu.cn (T.Z.)

<sup>2</sup> Department of Geography & Earth System Science, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium

<sup>3</sup> School of Environment and Spatial Informatics, China University of Mining and Technology, Daxue Road 1, Xuzhou 221116, China; weiqiang.liu@cumt.edu.cn

<sup>4</sup> School of Humanities and Law, Jiangsu Ocean University, Cangwu Road 59, Lianyungang 222005, China; saihu@jou.edu.cn

<sup>5</sup> Department of Research and Development, Shanghai Gongjing Environmental Protection Co., Ltd., Yuanjiang Road 525, Shanghai 201100, China; yanglh3@163.com

\* Correspondence: chenlq@cumt.edu.cn; Tel.: +86-516-8359-1327

**Abstract:** This study aims to integrate multisource data to model the relative soil moisture (RSM) over the Chinese Loess Plateau in 2017 by stepwise multilinear regression (SMLR) in order to improve the spatial coverage of our previously published RSM. First, 34 candidate variables (12 quantitative and 22 dummy variables) from the Moderate Resolution Imaging Spectroradiometer (MODIS) and topographic, soil properties, and meteorological data were preprocessed. Then, SMLR was applied to variables without multicollinearity to select statistically significant ( $p$ -value < 0.05) variables. After the accuracy assessment, monthly, seasonal, and annual spatial patterns of RSM were mapped at 500 m resolution and evaluated. The results indicate that there was a high potential of SMLR to model RSM with the desired accuracy (best fit of the model with Pearson's  $r = 0.969$ , root mean square error = 0.761%, and mean absolute error = 0.576%) over the Chinese Loess Plateau. The variables of elevation (0–500 m and 2000–2500 m), precipitation, soil texture of loam, and nighttime land surface temperature can continuously be used in the regression models for all seasons. Including dummy variables improved the model fit both in calibration and validation. Moreover, the SMLR-modeled RSM achieved better spatial coverage than that of the reference RSM for almost all periods. This is a significant finding as the SMLR method supports the use of multisource data to complement and/or replace coarse resolution satellite imagery in the estimation of RSM.

**Keywords:** relative soil moisture; Chinese Loess Plateau; stepwise multilinear regression; dummy variables



**Citation:** Yuan, L.; Li, L.; Zhang, T.; Chen, L.; Liu, W.; Hu, S.; Yang, L. Modeling Soil Moisture from Multisource Data by Stepwise Multilinear Regression: An Application to the Chinese Loess Plateau. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 233. <https://doi.org/10.3390/ijgi10040233>

Academic Editor: Wolfgang Kainz

Received: 26 January 2021

Accepted: 4 April 2021

Published: 6 April 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Soil moisture (SM) is widely recognized as a vital land surface variable that associates with land–atmosphere interaction [1,2], rainfall–runoff processes [3], water–energy balance [4], and climate change [5]. Accurate characterization of SM is beneficial for applications such as weather and climate modeling, agricultural and water resources management over larger spatial extents [6]. Accurate and timely SM estimation at a relevant spatiotemporal scale is a sound strategy to forecast droughts/floods [7,8] and schedule irrigation [9,10] for the sustainability and productivity of agriculture, particularly in arid and semi-arid regions like the Chinese Loess Plateau (CLP).

Various remote sensing data, spanning almost all regions of the electromagnetic spectrum from bands of microwave to visible, have been utilized for SM retrieval since

the 1970s [11–14]. In order to periodically obtain spatiotemporal maps of the SM, a variety of methods and techniques of SM estimation have been proposed as well [6]. However, microwave sensors (active and passive) cannot monitor SM well at local and regional scales due to their coarse spatial resolutions (a few tens of kilometers) [12,15,16]. Synthetic aperture radar (SAR) systems offer a better spatial resolution for SM retrievals but with a long revisit time. In addition, SM retrievals from SAR and microwave sensors are greatly affected by soil surface roughness, vegetation cover, and other relevant factors [17]. Numerical simulations, involving the use of land surface features retrieved by visible/near-infrared/thermal-infrared bands (e.g., vegetation, land surface temperature (LST), and surface albedo), have long been the primary methods for obtaining large-scale SM. These methods are based on Moderate Resolution Imaging Spectroradiometer (MODIS) data and land surface models [2,18,19]. However, the main problem associated with this method is that optical sensors cannot penetrate clouds and vegetation, which highly influences the quality of the SM estimation results [18,20].

One approach for obtaining accurate estimates of high-resolution SM is the disaggregation of microwave-derived SM using high-resolution disaggregated data like thermal-infrared and visible/near-infrared data [15,21–23]. The issue of this kind of downscaling method is the evaluation and validation procedure because of the impact of the uncertainties in input data and the scarcity of ground data [24]. Apart from the synergy of different data [25–28], another approach is to combine different estimation methods for SM retrieval [11,29–31]. Although Yuan et al. improved SM estimation concerning the spatiotemporal coverage for the CLP using MODIS-derived apparent thermal inertia (ATI) and Temperature Vegetation Dryness Index (TVDI), certain pixels did not have SM values. This led to the display of incomplete SM maps at the monthly scale [32].

In addition, SM variability is affected by a variety of environmental factors: soil properties (e.g., soil texture and organic matter) [33,34] at the plot scale, topography [35] and land cover [36,37] at the local scale, and precipitation, evapotranspiration [38,39], relative humidity, and temperature [40] obtained from meteorological data at the regional scale [41,42]. An effective approach to SM modeling with high accuracy and spatial resolution should integrate remotely sensed surface information, regional meteorological data, topographic data, and data of soil properties [18]. Data-driven methods for the estimation of SM include multivariate analyses, data assimilation, and machine learning techniques [43–45]. An SM data assimilation scheme is to simulate dynamic SM at spatiotemporal scales using estimated soil parameters and weather forcing based on a hydrological model [46,47]. The machine learning technique is computationally intensive (e.g., random forest (RF) [48,49], artificial neural network (ANN) [50–53], support vector regression (SVR) [54–56], and regression trees (RT) [57,58]) and used to build mathematical models based on training sets and covariates to extract SM information from the available data [59,60].

Multivariate analyses, especially regression analysis (multilinear regression (MLR) [61], stepwise multilinear regression (SMLR) [53,62], Gaussian process regression [63,64], partial linear squares regression [65], and sparse multiple linear regression [66]) are widely used to model SM. Among these methods, MLR, the most basic form of linear regression, predicts a single dependent variable from multiple independent variables. Yang et al. [67] performed an MLR model to estimate SM based on observed environmental variables (i.e., land use, topography, and meteorology) in the Danangou catchment (3.5 km<sup>2</sup>) of the CLP. The authors used the variable of land use and terrain indexes as dummy variables (i.e., a numeric variable that represents categorical data) to build the model [68], and concluded that the SMLR model was the most effective and economical among models [69]. SMLR, which basically repeats MLR many times, is a method of regressing multiple variables while removing those that are not significant [70]. Variables including the day of the year, canopy height, and NDVI were calibrated with SMLR and ANN to estimate fuel moisture content in tallgrass prairie [53]. Moreover, categorical variables like land cover [36], elevation [18,35], slope gradient, slope aspect, and soil texture [18,34,37,40] also be applied for modeling SM. The SMLR method was also used to predict soil water infiltration in a dry flood plain

of eastern Iran [71], to estimate the plant available water content of unsaturated soil [70]. When the variable of micro-porosity was included to estimate the availability of water in the soil, the SMLR (which features computation efficiency and ease of interpretation) was simpler to use compared with both ANN and RT [58].

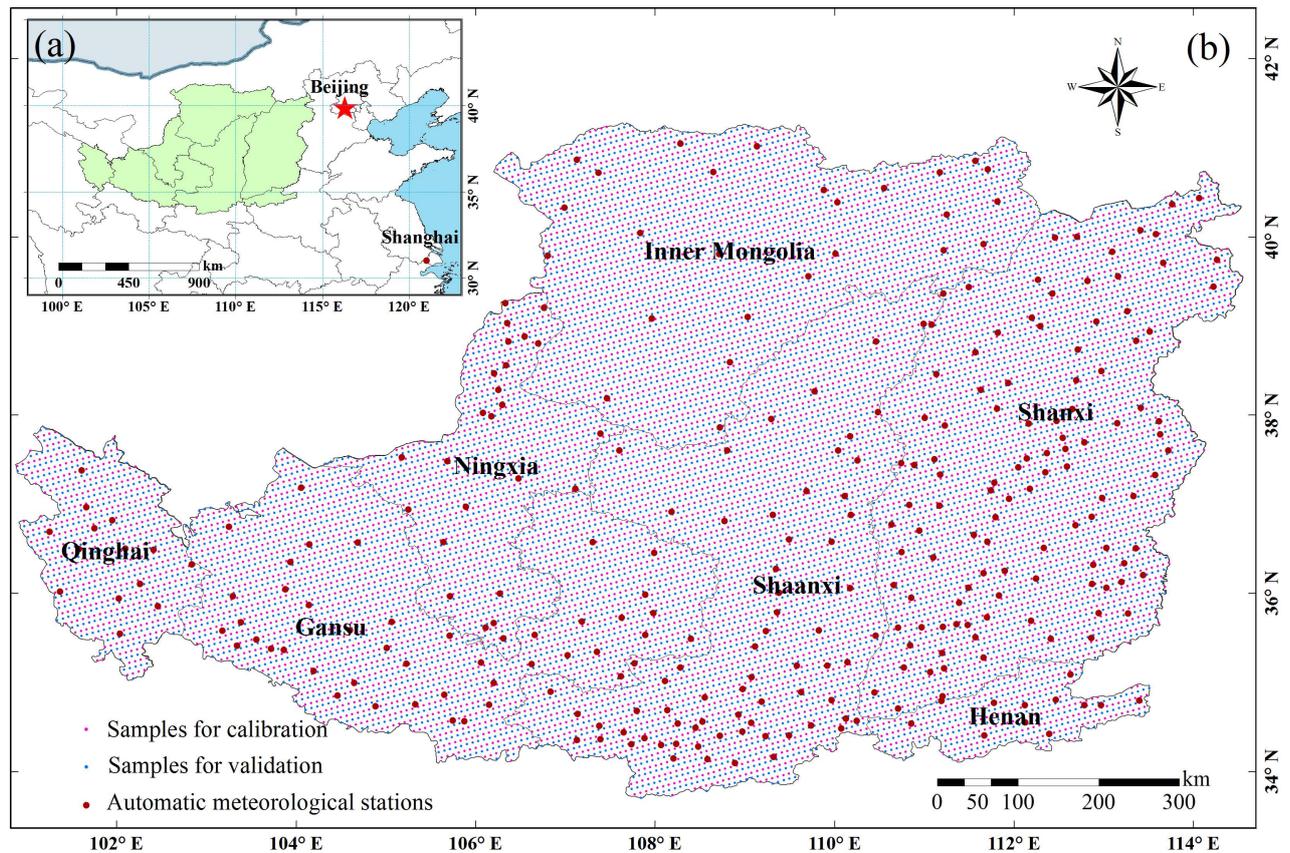
However, from the perspective of statistical analysis, the problem is more challenging if the sample size is not sufficient to apply the SMLR model [72,73]. Thus, the sparse in situ SM observations (e.g., only 49 available annual SM observations) are not suitable to be regarded as the reference SM for the entire CLP (640,000 km<sup>2</sup>) [32]. The incomplete reference SM maps from our previously published, to some extent, were still not available for certain applications. Thus, using the SMLR method over the CLP to estimate SM makes the best possible use of all ancillary data (translating the available data into the required data), particularly data that are relatively inexpensive and easily accessible [58].

The aim of this study is to integrate multisource data (MODIS and topographic, soil properties, and meteorological data) to model the relative soil moisture (RSM) at a spatial resolution of 500 m over the Chinese Loess Plateau in 2017 by stepwise multilinear regression (SMLR) in order to improve the spatial coverage of our previously published RSM. The previously published RSM was produced using MODIS-derived ATI and TVDI and is regarded as the reference RSM data in the present study. Detailed explanations for generating reference RSM are provided in Section 3. First, 34 candidate variables (12 quantitative and 22 dummy variables) were preprocessed. Then, SMLR was applied to variables without multicollinearity to select statistically significant ( $p$ -value < 0.05) variables. The regression models and the accuracy of the modeled RSM were evaluated at monthly, seasonal, and annual scales. Finally, the modeled RSM was analyzed to better understand the spatiotemporal characteristics of the RSM.

## 2. Study Area

This study area is the CLP—Chinese Loess Plateau (100°54′–114°33′ E and 33°43′–41°16′ N), northwestern China, which covers an area of approximately 640,000 km<sup>2</sup>, spanning seven provinces (Figure 1a). The landscape is strongly shaped by wind–water erosion and has a highly fractured landform of gullies [74]. Both the mean annual temperature and precipitation gradually decrease from the southeast (14 °C and 750 mm) to the northwest (4 °C and 200 mm) [75]. The CLP has been categorized as the most seriously eroded landscape in the world because of its loose and erodible soil [76,77]. The already low and concentrated precipitation (the mean annual precipitation is 420 mm and 55–78% fall in the wet season from July to September) makes the CLP particularly vulnerable to drought [78]. In addition, SM over the CLP shows significant spatial variation due to climatic characteristics and fragmented topography [36,79]. Therefore, quantitative estimations of the SM over the CLP are more important than for other regions.

In the present study, to ensure uniform distribution of calibration and validation samples (without overlapping), both the 7814 calibration samples (small red points in Figure 1b) and the 7824 validation samples (small blue points in Figure 1b) were selected at 10 km intervals while the distance between the adjacent calibration and validation samples was 5 km. A total of 298 Chinese automatic meteorological stations (large red points in Figure 1b) provided hourly precipitation and relative air humidity observation data over the CLP.



**Figure 1.** Study area: (a) Location of the Chinese Loess Plateau (CLP) in China; (b) spatial distribution of samples for both calibration and validation (both sampled at 10 km intervals) and 298 automatic meteorological stations used in the study. The distance between adjacent calibration and validation samples was 5 km.

### 3. Materials and Methods

#### 3.1. Soil Moisture Data

The relative soil moisture (RSM) represents the percentage of SM that accounts for the moisture storage capacity and was used to describe the SM levels in the present study. The monthly, seasonal, and annual RSM maps of the CLP in 2017 (previously published [32]) were used and regarded as the reference RSM for modeling in this study. The previously published RSM was generated via 8-day RSM maps. The overall 8-day RSM was combined at a 500 m resolution by corresponding subregional RSM, which was produced with three groups of selected optimal NDVI thresholds using MODIS-derived ATI (apparent thermal inertia) and TVDI (temperature vegetation dryness index), and the average of ATI and TVDI against 20 cm depth in situ RSM observations [32]. Here, many studies pointed out TVDI and ATI could adequately reflect the changes of RSM at a 20 cm depth [11,80–82]. In terms of RSM estimation using the ATI-based model, soil thermal inertia (TI) is described as a thermal property of soil that characterizes its resistance to temperature change and has been used for near-surface SM retrieval [11,83]. ATI, to simplify the TI, is calculated by spectral surface albedo and the diurnal land surface temperature (LST) range [31,84]. The ATI method has been used for monitoring RSM for bare soil or sparsely vegetated regions. In addition, as an effective method based on the NDVI-LST feature space, the TVDI-based model considers vegetation coverage in RSM estimation and has been widely applied to vegetated areas. To estimate 8-day subregional RSM, the overall CLP was divided into three subregions (the ATI subregion, the TVDI subregion, and the ATI/TVDI subregion) according to the NDVI of individual pixels. The ATI-based model, the TVDI-based model, and the ATI/TVDI joint model were used in the ATI subregion, the TVDI subregion, and

the ATI/TVDI subregion, respectively, and corresponding subregional RSM data were obtained. Therefore, the equations were used as follows:

$$RSM_{overall} = \begin{cases} RSM_{ATI} = a_{ATI} \times ATI + b_{ATI} & NDVI \in [0, NDVI_{ATI}] \\ RSM_{ATI/TVDI} = a_{ATI/TVDI} \times \frac{ATI+TVDI}{2} + b_{ATI/TVDI} & NDVI \in (NDVI_{ATI}, NDVI_{TVDI}] \\ RSM_{TVDI} = a_{TVDI} \times TVDI + b_{TVDI} & NDVI \in (NDVI_{TVDI}, 1] \end{cases} \quad (1)$$

where  $RSM_{overall}$  represents the overall RSM and it is combined by three subregional RSM ( $RSM_{ATI}$ ,  $RSM_{ATI/TVDI}$ , and  $RSM_{TVDI}$ ).  $RSM_{ATI}$  and  $RSM_{TVDI}$  are the RSM estimated by the ATI-based and TVDI-based models, respectively, and  $RSM_{ATI/TVDI}$  is the RSM estimated by the ATI/TVDI joint model.  $a_{ATI}$  and  $b_{ATI}$  are coefficients from fitting the ATI values and in situ RSM observations in the ATI subregion.  $a_{TVDI}$  and  $b_{TVDI}$  are coefficients from fitting the TVDI values and in situ RSM observations in the TVDI subregion.  $a_{ATI/TVDI}$  and  $b_{ATI/TVDI}$  are coefficients from fitting the average value of ATI and TVDI and in situ RSM observations in the ATI/TVDI subregion.  $NDVI_{ATI}$  and  $NDVI_{TVDI}$  are the selected optimal thresholds for generating subregions.

Three optimal NDVI thresholds ( $NDVI_0$  was used for computing TVDI, and both  $NDVI_{ATI}$  and  $NDVI_{TVDI}$  for dividing the entire CLP) were identified with the best validation results of subregions for 8-day periods. To assess the performance of the models over the CLP, the Pearson's  $r$  and the mean absolute error (MAE) of the reference RSM against in situ RSM observations were calculated (at the monthly, seasonal, and annual scales). The  $r$  and MAE varied from 0.47 in October to 0.68 in January and from 3.02% in March to 4.97% in October, respectively, on the monthly scale (Table A1) [32]. Since the  $r$  of reference RSM and in situ RSM observations had a moderate correlation coefficient ( $r = 0.73$  for the annual scale) [85], the reference RSM was assumed to be the actual RSM with larger spatial coverage than those in situ RSM observations and the reference RSM kept a better trend with the in situ observed RSM at the station scale. The area of the reference RSM within five months (April, May, July, August, and October) was less than half of the entire study area ( $\sim 32 \times 10^4 \text{ km}^2$ ) (Table A1). Thus, the SMLR method was applied to improve the spatial coverage of our previously published RSM based on the reference RSM and multisource data.

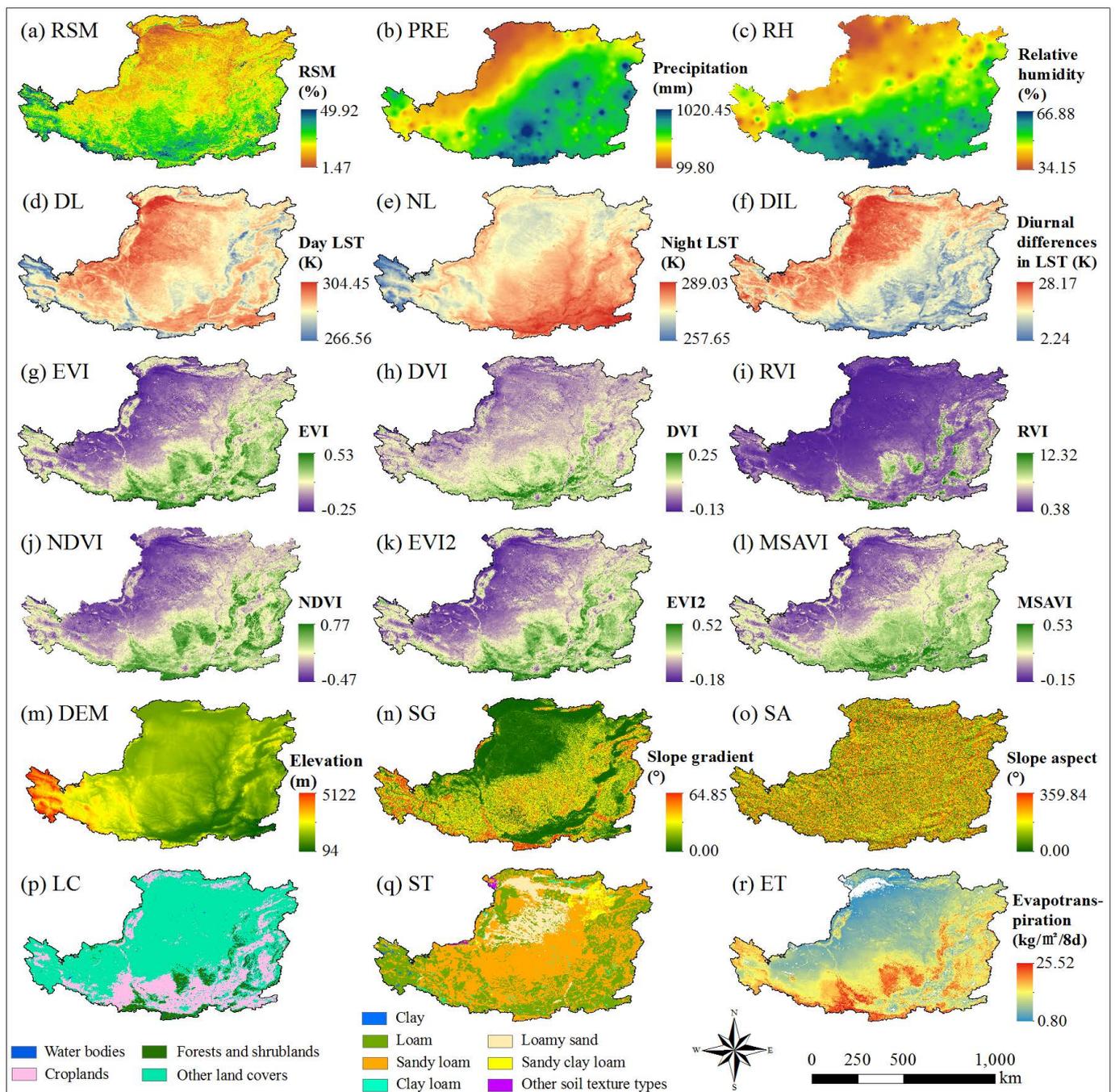
### 3.2. Candidate Variables

A total of 17 features (12 quantitative and five categorical variables), including MODIS-derived features, topographical features, soil properties, and meteorological features, were selected as candidate variables for RSM modeling using SMLR. These candidate variables in our study were applied to estimate RSM in previous studies [18,34,39,40]. To be more specific, the 12 quantitative variables included daytime LST (DL) [18,40], nighttime LST (NL), diurnal differences in LST (DIL), evapotranspiration (ET) [18,38,39], enhanced vegetation index (EVI), difference vegetation index (DVI), ratio vegetation index (RVI), normalized difference vegetation index (NDVI) [34], enhanced vegetation index 2 (EVI2), modified soil-adjusted vegetation index (MSAVI), precipitation (PRE) [38,39], and relative humidity (RH) [86]. Moreover, the categorical variables were land cover (LC) [36], elevation (DEM) [18,35], slope gradient (SG), slope aspect (SA), and soil texture (ST) [18,34,37,40], which were declared as dummy variables in the regression equations. The variables used in this study are described in Table 1 (see Table A2 for the calculation of the vegetation indexes).

**Table 1.** Variables used in this study (green color rows represent dummy variables used in the study).

Sources (Types)	Products	Parameters	Variables	Abbr.	Units	Spatial/Temporal Resolution
MODIS	MOD11A2 in 2017	Daytime/nighttime LST	Day LST Night LST Diurnal differences in LST	DL NL DIL	K	1 km, 8-day
	MOD16A2 in 2017	Evapotranspiration	Evapotranspiration	ET	kg/m <sup>2</sup> /8 d	500 m, 8-day
	MOD09A1 in 2017	Surface reflectance	Enhanced vegetation index Difference vegetation index Ratio vegetation index Normalized difference vegetation index Enhanced vegetation index 2 Modified soil-adjusted vegetation index	EVI DVI RVI NDVI EVI2 MSAVI	No unit	500 m, 8-day
	MCD12Q1.Type2 in 2017	Land cover	Land cover	LC	No unit	500 m, 1 year
Topographic data	SRTM DEM	DEM	Elevation	DEM	m	
	SRTM SLOPE	Slope	Slope gradient	SG	°	90 m, N/A
	SRTM ASPECT	Aspect	Slope aspect	SA	°	
Soil properties	Sand/silt/clay	Soil texture	Soil texture	ST	No unit	1:1,000,000, N/A
Meteorological data	Hourly observation data in 2017	Precipitation	Precipitation	PRE	mm	N/A, hourly
		Relative humidity	Relative humidity	RH	%	

A dummy variable is an artificial variable that is created to represent an attribute with two or more distinct levels/categories [87,88]. To avoid the dummy variable trap (a scenario in which independent variables are collinear), one less dummy variable (n-1) than the categorical values (n) was used [89]. Thus, dummy variables, including elevation (seven categories and six dummy variables), land cover (three categories and two dummy variables), slope aspect (four categories and three dummy variables), slope gradient (six categories and five dummy variables), and soil texture (seven categories and six dummy variables) were used in this study. The spatial pattern of the annual reference RSM and 17 features are shown in Figure 2. Detailed pre-processing from features to candidate variables is given in the later section.



**Figure 2.** Distribution of annual reference relative soil moisture (RSM) (a) and 17 features (12 quantitative and five categorical variables) (b–r).

### 3.2.1. MODIS Data

The LST and vegetation indexes are two important parameters closely related to RSM and are often used on RSM estimations [90]. The 8-day 1 km composite LST product (MOD11A2), the 8-day 500 m evapotranspiration (ET) product (MOD16A2), the 8-day 500 m surface reflectance product (MOD09A1), and the annual 500 m land cover product (MCD12Q1. Type2) of 2017 were used to develop the variables of DL, NL, ET, LC, and six vegetation indexes. The utilized MODIS data were downloaded from the website of the level-1 and Atmosphere Archive and Distribution System (LAADS) Distributed Archive Center (DAAC) [91]. In order to facilitate the utilization of MODIS products, all pixels were processed for quality assurance based on the valid range before further computing. The

LST products, including DL and NL, were resampled to a resolution of 500 m (reference RSM resolution) and were chosen to calculate the diurnal differences in LST ( $DIL = DL - NL$ ) [18,40]. Pixels were masked when the DL was less than 0 °C, suggesting snow coverage or frozen soil at that time and location. Strong positive relationships between RSM and the vegetation index (e.g., NDVI) were found for Australia [92], tallgrass prairies in the USA [93,94], croplands in North China [95], and East Africa [96]. Surface reflectance products (including seven bands) were chosen to calculate the five vegetation indices: EVI, DVI, RVI, NDVI, EVI2, and MSAVI [34]. Monthly, seasonal, and annual data applied for modeling were combined by these 8-day composite products.

The influences of land covers on RSM are complex [36,37]. According to the MCD12Q1. Type2 from the University of Maryland (UMD) land cover classification scheme (with 16 different cover types) [97], the land cover data were reclassified here into three categories, namely croplands, forest, and shrublands, as well as other land covers in the study area (Table 2). With regard to the variable of land cover (a categorical variable), with other land covers as reference (row in the orange color background in Table 2), the remaining two land cover types (LC1 and LC2, rows in the green color background in Table 2) were integrated as dummy variables in the modeling procedure (with only two values; 0 and 1) [98,99].

**Table 2.** Land cover reclassification of the University of Maryland (UMD) classification system (green color rows represent dummy variables of LC1 and LC2 and reference land covers are the rows with the orange color background).

Class	UMD Classification	Proportion (%)	Regroup	Variables
0	Water	0.163	-	-
1	Evergreen needle leaf forest	6.388	Forest and shrublands	LC2
2	Evergreen broad leaf forest			
3	Deciduous needle leaf forest			
4	Deciduous broad leaf forest			
5	Mixed forest			
6	Closed shrublands			
7	Open shrublands			
8	Woody savannas	69.180	Other land covers	
9	Savannas			
10	Grasslands			
13	Urban and built-up			
15	Barren or sparsely vegetated	24.180	Croplands	LC1
12	Croplands			
14	Cropland/natural vegetation mosaic			
11	Permanent wetlands	0.087	-	-
255	Unclassified	0.002	-	-

### 3.2.2. Topographic Data

Digital elevation model (DEM) data were used because the distribution of the RSM and other features are directly related to elevations [18,100]. The 90 m spatial resolution Shuttle Radar Topography Mission (SRTM) DEM, SG, and SA datasets for the study area were downloaded from the Geospatial Data Cloud (GDC) Platform (<http://www.gscloud.cn/>, accessed on 11 October 2020) and were then resampled to a 500 m spatial resolution. These three chosen features were used as dummy variables in this study. DEM was classified into seven categories at 500 m intervals. With elevations exceeding 3000 m as the reference, the remaining six elevation categories were integrated as dummy variables (DEM1, DEM2, DEM3, DEM4, DEM5, and DEM6) in the modeling procedure (Table 3). A total of 52% of the CLP was in the elevation categories from 1000 to 1500 m. Slope gradients, ranging from 0° to 64.85° over the CLP (Figure 2n), were classified into six categories at 5° intervals. With SG exceeding 25° as the reference, the remaining five SG categories were also integrated as dummy variables (SG1, SG2, SG3, SG4, and SG5) in the modeling procedure (Table 3).

**Table 3.** Elevation and slope gradient classification for the utilized dummy variables (green color rows represent dummy variables of DEM1, DEM2, DEM3, DEM4, DEM5, DEM6, SG1, SG2, SG3, SG4, and SG5, and reference elevation (3000 m and above) and reference slope gradient (25° and above) are the rows with the orange color background).

Type	Class	Range	Proportion (%)	Variables
Elevation	1	0–500m	4.566	DEM1
	2	500–1000m	12.727	DEM2
	3	1000–1500m	52.080	DEM3
	4	1500–2000m	18.736	DEM4
	5	2000–2500m	6.480	DEM5
	6	2500–3000m	2.383	DEM6
	7	3000 m and above	3.028	
Slope gradient	1	0–5°	43.450	SG1
	2	5–10°	18.171	SG2
	3	10–15°	16.800	SG3
	4	15–20°	11.514	SG4
	5	20–25°	5.889	SG5
	6	25° and above	4.176	

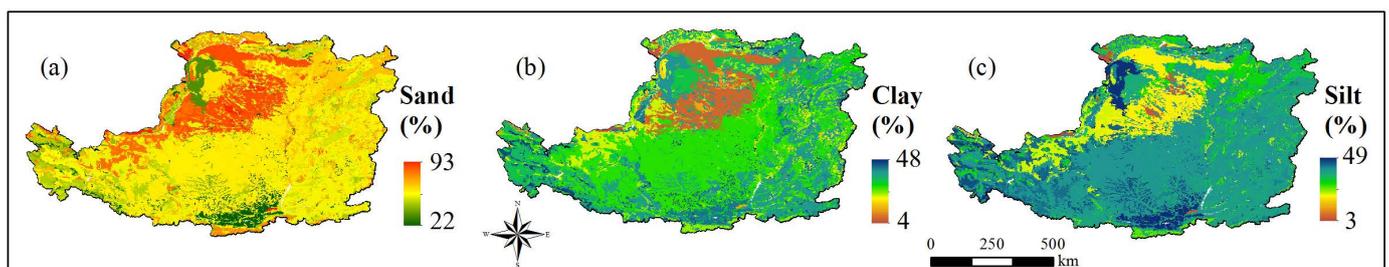
The slope aspect (SA) played an important role in the distribution of RSM at the hillslope domain and should be considered when attempting to characterize RSM variability in gullied regions [101]. Here, SA was divided into four categories (semi-shady, semi-sunny, shady, and sunny, as shown in Figure A1). Using the sunny category as the reference, the remaining three categories were integrated as dummy variables (SA1, SA2, and SA3) in the modeling procedure (Table 4).

**Table 4.** Slope aspect classification of dummy variables (green color rows represent dummy variables of SA1, SA2, and SA3 and the reference slope aspect of sunny is the row with the orange color background).

Class	Range (°)	Directions	Proportion (%)	Variables
1	45–90, 270–315	Semi-shady	26.180	SA1
2	90–135, 225–270	Semi-sunny	25.109	SA2
3	0–45, 315–360	Shady	24.177	SA3
4	135–225	Sunny	24.534	

### 3.2.3. Soil Properties Data

RSM was strongly influenced by soil texture (ST) [102,103]. Maps of the sand, clay, and silt content for the study area were provided by the Data Center for Resource and Environmental Sciences, Chinese Academy of Sciences (RESDC, <http://www.resdc.cn/>, accessed on 20 August 2020) (Figure 3). High percentage sand content was clustered in the northwestern region of the CLP, and low percentage sand content and high percentage silt content covered the southern CLP.



**Figure 3.** Spatial distribution of sand (a), clay (b), and silt (c) over the CLP.

Soil texture was classified into 12 types using the U.S. Department of Agriculture (USDA) Soil Texture Classification System (Figure A2 and Table 5). A total of seven textural classes were identified over the CLP and more than half of the CLP (51.493%) was covered by sandy loam. The textural classes were integrated as dummy variables (ST1, ST2, ST3, ST4, ST5, and ST6) using the class of sandy loam as the reference group in the modeling procedure (Table 5).

**Table 5.** Soil texture classification for dummy variables (green color rows represent dummy variables of ST1, ST2, ST3, ST4, ST5, and ST6, and reference soil texture of sandy loam is the row with the orange color background).

Class	Soil Texture	Proportion (%)	Variables
1	Silt	-	-
2	Silt loam	-	-
3	Loam	34.292	ST1
4	Silty clay loam	-	-
5	Clay loam	1.088	ST2
6	Silty clay	-	-
7	Clay	0.206	ST3
8	Sandy clay loam	2.531	ST4
9	Sandy clay	-	-
10	Loamy sand	10.017	ST5
11	Sand	0.373	ST6
12	Sandy loam	51.493	

### 3.2.4. Meteorological Data

Precipitation (PRE), both the amount and intensity, has been shown to be a major driver of SM dynamics [104–106]. PRE and relative humidity (RH) were also applied to RSM estimation [86]. The network of in situ automatic meteorological stations of the Chinese Meteorological Data Service Center (CMDC) provided the meteorological inputs required by the model: RH and PRE (both acting as quantitative variables). Hourly RH (%) and PRE (mm) were recorded at 298 automatic meteorological stations (Figure 1) over the CLP in 2017 [107]. Concerning RH, for an accurate temporal match between in situ observation data and monthly reference RSM (produced by averaging corresponding 8-day RSM), the daily granule acquisition time of MOD09GA products was obtained to serve as the reference for selecting corresponding in situ RH observations. The monthly in situ RH value at each meteorological station was computed by averaging the daily RH. Then, the spatially interpolated inverse distance weighted (IDW) method was applied to express the RH spatial patterns at a resolution of 500 m [108]. The same procedure was performed for seasonal and annual RH maps (Figure 2c). The mean annual RH over the CLP in 2017 ranged from 34.15% to 66.88%.

Regarding PRE, daily PRE with cumulative values were also rescaled to monthly, seasonal, and annual temporal resolutions. After interpolation by inverse distance weighting, monthly, seasonal, and annual PRE maps were generated [18,109]. The maximum and minimum total annual PRE in 2017 was 1020.15 mm and 99.80 mm, respectively (Figure 2b). Both PRE and RH appeared to be gradually decreasing from southeast to northwest and the year 2017 was a typical year from a climate perspective [110,111].

### 3.3. Stepwise Multilinear Regression Modeling

The calibration samples for each period were used to constructing stepwise multilinear regression (SMLR) models first. As we presented in Figure 1, 7814 calibration samples and 7824 validation samples were selected at 10 km intervals while the distance between the adjacent calibration and validation samples was 5 km. Table 6 displays the different used samples for calibration and validation due to different RSM areas for each period. The next step is to determine the actual set of variables used from 34 candidate variables (12 quantitative variables and 22 dummy variables) in the final regression. SMLR is

routinely used for finding important variables while multicollinearity among variables often undermines its performance [98,112]. To select variables without multicollinearity, the general MLR model was first applied using calibration samples with 34 independent variables for each period (monthly, seasonal, and annual data) (Figure 4). The MLR model involves a series of single factor correction coefficients and, in this study, does not consider interactions (excluding transformed variables) among variables. An MLR can be represented as:

$$\gamma = \beta_0 + \sum_{i=1}^N \beta_i X_i + \varepsilon \quad (2)$$

where  $\gamma$  represents the dependent variable (reference RSM), and  $\beta_0$  and  $\beta_i$  represent the constant offset and regression coefficients of the corresponding explanatory variables  $X_i$ , respectively. The deviation between model outputs and reference RSM represents the model bias  $\varepsilon$ .

**Table 6.** The number of samples for calibration and validation at the seasonal and annual scales.

Period	In Calibration		In Validation	
	No. of Samples	Proportion <sup>1</sup> (%)	No. of Samples	Proportion <sup>1</sup> (%)
Annual	7484	0.241	7493	0.242
Winter	6771	0.255	6763	0.254
Spring	6891	0.222	6879	0.222
Summer	6419	0.214	6415	0.214
Autumn	7415	0.248	7418	0.248

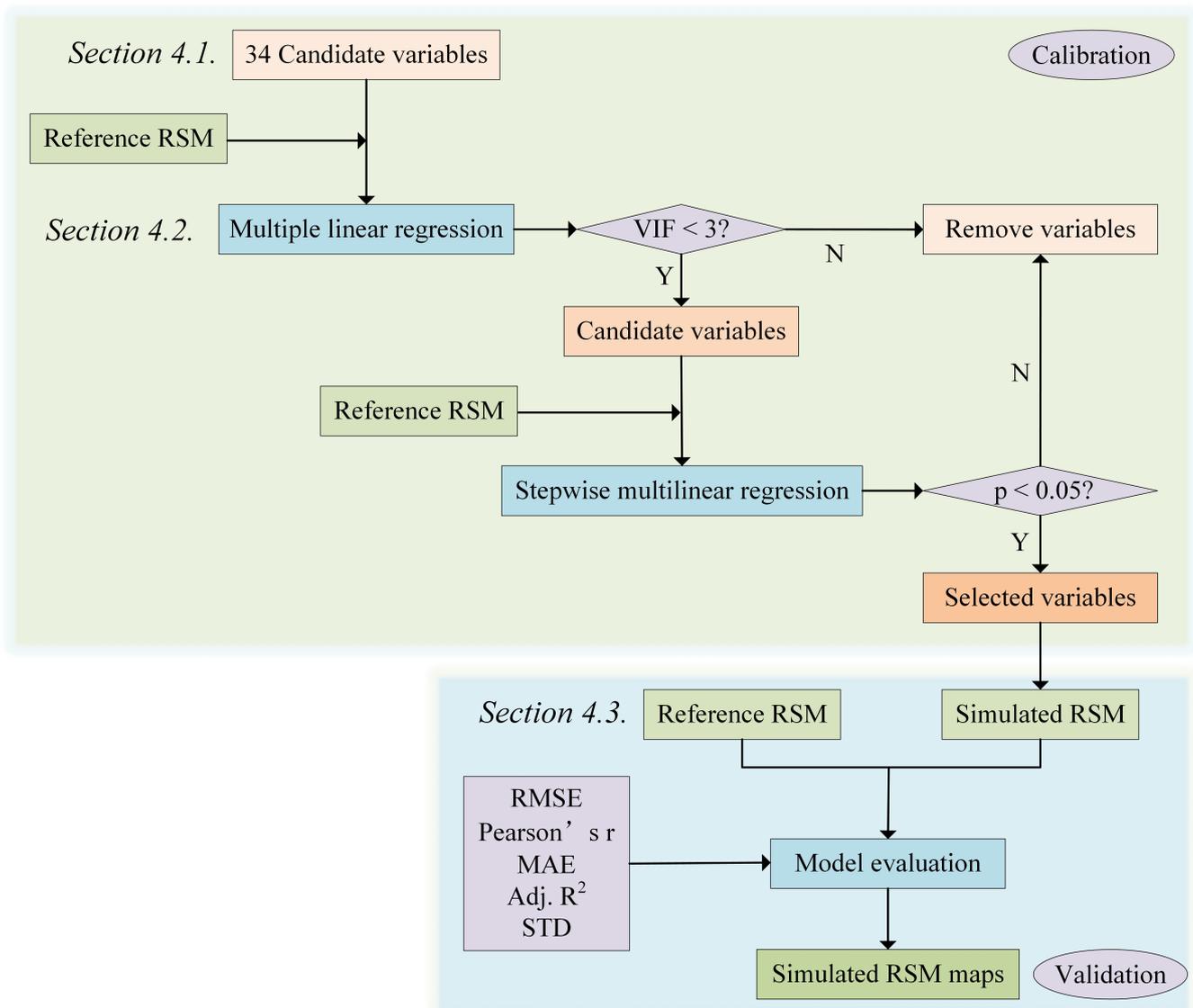
<sup>1</sup> The proportion is the percentage of sample size for calibration or validation in total valued pixels for each period.

Since collinearity likely exists among 34 candidate variables, the Variation Inflation Factor (VIF) [98] was used to examine it:

$$\text{VIF} = \frac{1}{1 - R_i^2} \quad (3)$$

where  $R_i$  represents the correlation coefficient between the  $i^{\text{th}}$  predictive variable and the remaining predictive variables. No multicollinearity exists if VIF is less than 3 [109,113].

If VIF exceeded 3 (which indicates multicollinearity), the variable with the highest VIF was removed and the model was re-evaluated. Then, the candidate variables with a VIF of less than 3 were prepared for SMLR modeling. In an SMLR analysis, the most significant or least significant variable is iteratively added to or removed from the MLR model based on its statistical significance [98]. Statistically significant variables were identified through continual regression iterations in the linear regression equation ( $p < 0.05$  was applied in this study). This method can effectively select powerful features for the construction of a good predictive model and has been widely used in different fields [89,98,99,114], including SM estimation [34,62]. In addition, although there are many different strategies for selecting variables for a regression model, all possible regression procedures should be used if there are no more than fifteen candidate variables and the SMLR could be used for more than fifteen candidate variables [115]. As such, the SMLR was considered more suitable for constructing RSM models in this study.



**Figure 4.** Flowchart of the processing chain implemented for RSM modeling using stepwise multilinear regression. Each collection of boxes was labeled according to the corresponding sections in the main text.

### 3.4. Accuracy Assessment

The validation samples (presented in Figure 1b) for each period were used to assess the accuracy of the constructed SMLR models. Statistical metrics, including root mean square error (RMSE), Pearson's  $r$  ( $r$ ), Adjusted  $R^2$  (Adj.  $R^2$ ), MAE, and standard deviation (STD) were calculated to evaluate the performance of the simulation [11,116]. The agreement and degree of dispersion between the modeled RSM via SMLR and reference RSM were analyzed here in terms of these five classical statistical criteria for each period. The equations are presented in Table A3. The modeled RSM data, which used multisource data via the SMLR method at the monthly, seasonal, and annual scales, were evaluated and compared with the reference RSM.

## 4. Results

### 4.1. Stepwise Multilinear Regression Model

The SMLR models were established using selected variables to simulate RSM. The results of performing SMLR for constructing the RSM models are presented in Table 7. For the SMLR models, the model fit was generally assessed by its Adj.  $R^2$ . In this case,

the models that were developed to estimate RSM fitted best with the highest Adj.  $R^2$  of 0.912 (i.e., 91.2% of the variation of the dependent variable RSM can be explained by the change in the independent variables) and the lowest RMSE of 0.798% in winter among all four seasons in 2017. Moreover, the best model fitting results were obtained in December (Adj.  $R^2 = 0.938$  and RMSE = 0.753%) among the months (Table A4). The linear regression equation had the lowest Adj.  $R^2$  of 0.091 and highest RMSE of 4.182% in October. Focusing on intercepts of the models, Table 7 shows that there are both negative (in winter and autumn) and positive (in spring and summer) intercepts, and the annual regression model is identified with positive intercepts (125.231).

**Table 7.** Stepwise multilinear regression analysis of annual and seasonal RSM.

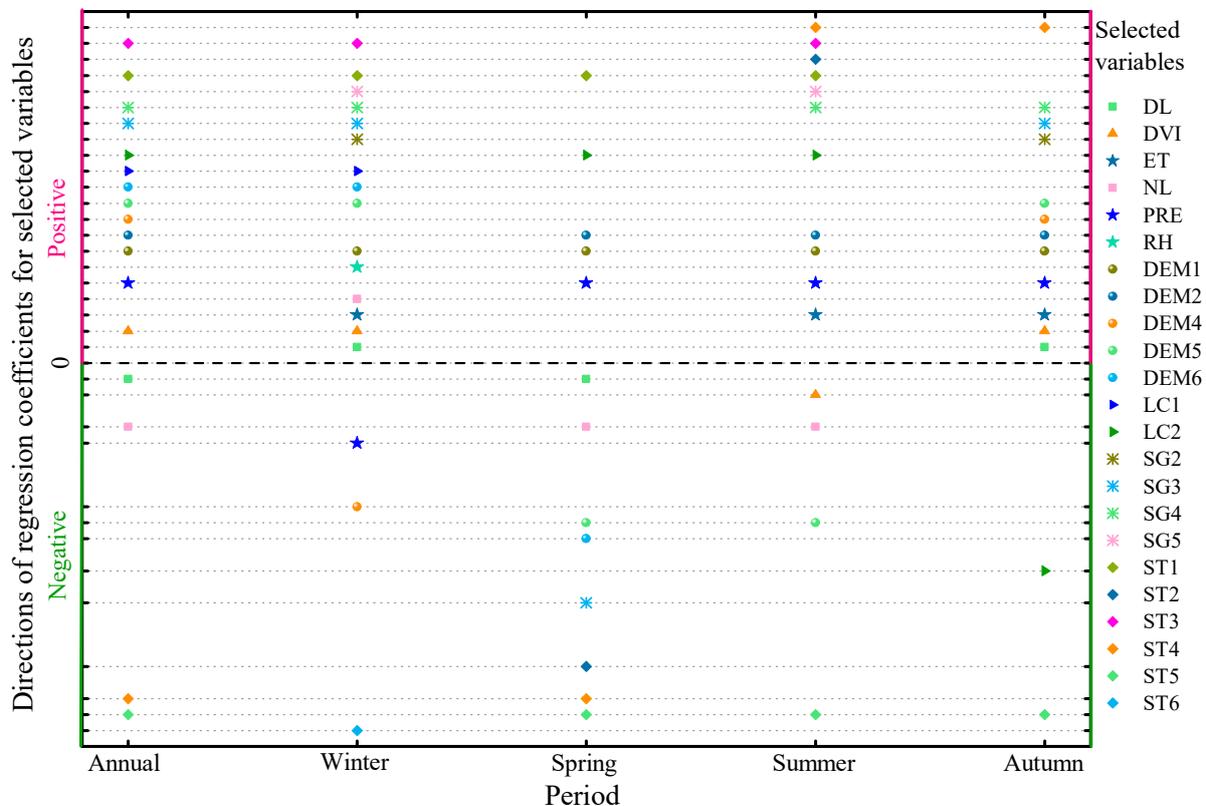
Period	Regression Model	Adj. $R^2$	RMSE	Max VIF
Annual	$RSM = 125.231 - 0.294*DL + 24.804*DVI - 0.114*NL + 0.004*PRE + 2.409*DEM1 + 0.884*DEM2 + 0.277*DEM4 + 0.552*DEM5 + 0.726*DEM6 + 0.632*LC1 + 0.398*LC2 + 0.144*SG3 + 0.163*SG4 + 0.237*ST1 + 1.826*ST3 - 1.607*ST4 - 0.721*ST5$	0.572	1.698	2.993
Winter	$RSM = -226.027 + 0.806*DL + 20.709*DVI + 0.032*ET + 0.030*NL - 0.014*PRE + 0.052*RH + 0.239*DEM2 - 0.263*DEM4 - 0.155*DEM5 + 0.185*DEM6 + 0.215*LC1 + 0.174*SG2 + 0.297*SG3 + 0.296*SG4 + 0.277*SG5 + 0.086*ST1 + 0.983*ST3 - 0.998*ST6$	0.912	0.798	2.195
Spring	$RSM = 195.013 - 0.564*DL - 0.059*NL + 0.019*PRE + 3.504*DEM1 + 1.750*DEM2 - 0.590*DEM5 - 0.760*DEM6 + 1.255*LC2 - 0.300*SG3 + 0.201*ST1 - 0.779*ST2 - 1.122*ST4 - 0.710*ST5$	0.540	2.978	2.428
Summer	$RSM = 193.779 - 6.122*DVI + 0.340*ET - 0.650*NL + 0.012*PRE + 2.152*DEM1 + 0.895*DEM2 - 0.728*DEM5 + 2.484*LC2 + 0.429*SG4 + 0.891*SG5 + 0.518*ST1 + 1.063*ST2 + 2.591*ST3 + 0.676*ST4 - 1.962*ST5$	0.592	3.834	2.242
Autumn	$RSM = -45.187 + 0.187*DL + 16.927*DVI + 0.285*ET + 0.006*PRE + 1.886*DEM1 + 1.095*DEM2 + 0.151*DEM4 + 0.375*DEM5 - 1.657*LC2 + 0.457*SG2 + 0.520*SG3 + 0.338*SG4 - 0.828*ST4 - 0.301*ST5$	0.283	2.202	2.366

Clearly, the selected variables and the number of selected variables for each regression model varied with the assessed periods. The directions of regression coefficients for variables in the SMLR models (Table 7) indicate the positive or negative correlations between RSM and the variables. The selected variables with directions of regression coefficients in the models at monthly, seasonal, and annual scales are shown in Figures 5 and A3. In total, 12 quantitative variables and five categorical variables (represented by 22 dummy variables) were used in this study. Among these 34 candidate variables, 27 variables were selected for monthly models.

Interestingly, the directions of the regression coefficients for an individual variable were not fixed; for example, the variable of PRE (labeled with blue stars in Figure A3). A positive correlation was found with the RSM in March, June, August, and November, but negative directions of regression coefficients were found in July, September, October, and December. Moreover, the selected variables of ST1, ST3, and DIL also kept the same direction in the regression models for the periods.

The number of selected variables for modeling at seasonal and annual scales (23 selected variables) was slightly lower than that at the monthly scale (27 selected variables). The variables of PRE with positive regression coefficients (except in winter), DEM1 (elevation of 0–500 m with positive regression coefficients) were used to simulate RSM in the seasonal and annual regression models. Focusing on the annual regression model, 17 selected variables of DVI, PRE, DEM1, DEM2, DEM4, DEM5, DEM6, LC1, LC2, SG3, SG4, ST1, and ST3 were positively correlated and DL, NL, ST4, ST5, were negatively correlated and were applied to simulate RSM. Among the five categorical variables (elevation, slope

gradient, slope aspect, land cover, and soil texture), four categorical variables were used (except for slope aspect) in the annual and seasonal linear equations.



**Figure 5.** Selected variables for modeling at seasonal and annual scales. Labels above and below 0 represent positive and negative directions of regression coefficients for selected variables in the regression model, respectively.

To identify the contribution of dummy variables to the SMLR models, SMLR was conducted without dummy variables at the seasonal and annual scales and Table 8 shows the comparison results. As expected, including dummy variables improved the model fit in the calibration as well as in the validation because almost all periods (except for summer) of the Adj. R<sup>2</sup> with dummy variables were higher than those without dummy variables. This might be contributed to the low accuracy of reference RSM at that time. It is also possible that other selected numerical variables, such as precipitation, have significant effects on RSM changes, while dummy variables appear to have relatively weak influence as a whole.

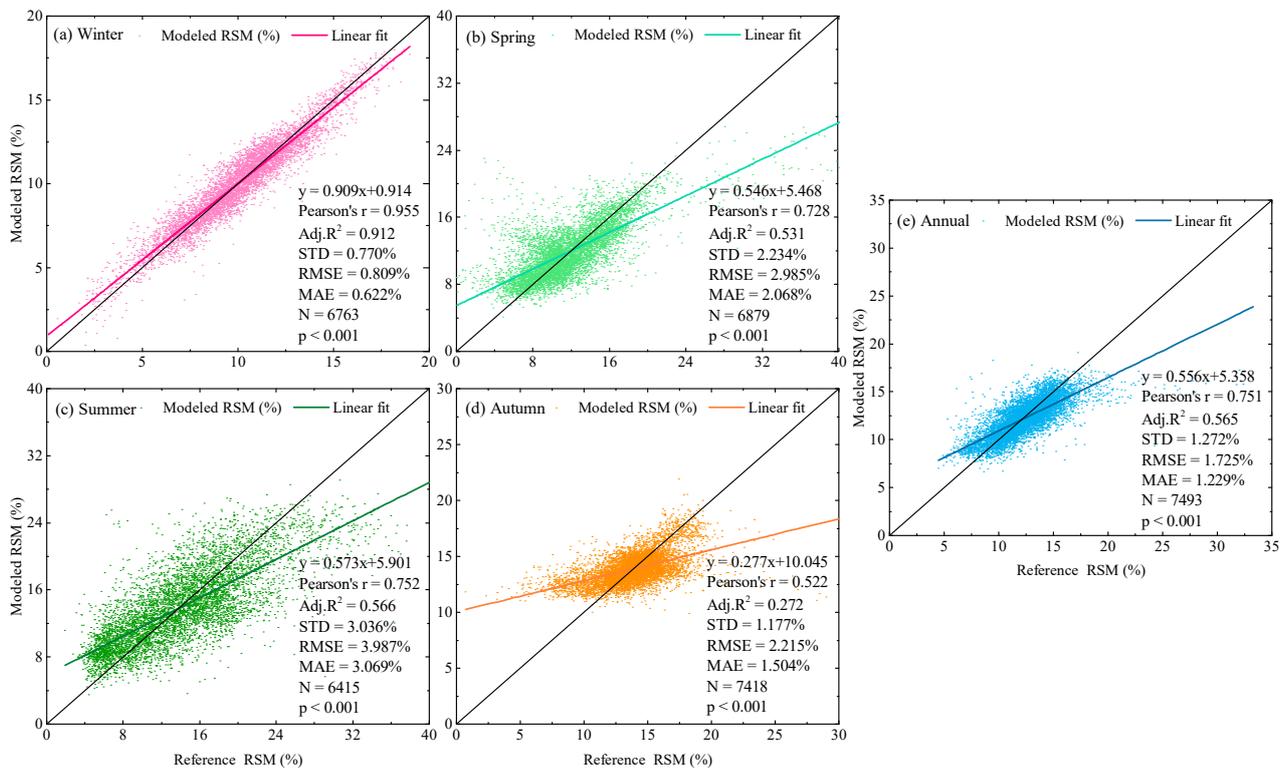
**Table 8.** Comparison of the accuracy of the stepwise multilinear regression (SMLR) models with and without dummy variables in calibration and validation at the seasonal and annual scales (the results of models with dummy variables that performed better are formatted in bold).

Period	In Calibration <sup>1</sup>		In Validation <sup>1</sup>	
	Adj. R <sup>2</sup> with Dummy Variables	Adj. R <sup>2</sup> without Dummy Variables	Adj. R <sup>2</sup> with Dummy Variables	Adj. R <sup>2</sup> without Dummy Variables
Annual	<b>0.572</b>	0.505	<b>0.565</b>	0.497
Winter	<b>0.912</b>	0.908	<b>0.912</b>	0.908
Spring	<b>0.540</b>	0.515	<b>0.531</b>	0.501
Summer	0.592	0.647	0.566	0.625
Autumn	<b>0.283</b>	0.227	<b>0.272</b>	0.215

<sup>1</sup> p-values < 0.05.

#### 4.2. Accuracy Assessment

Figure 6 illustrates the results of the assessment of RSM models at seasonal and annual time scales. The highest  $r$  of 0.955 and Adj.  $R^2$  of 0.912 and the lowest STD of 0.770%, RMSE of 0.809%, and MAE of 0.622% were found in winter (Figure 6a). The annual validation result was characterized by a relatively high correlation coefficient ( $r > 0.75$ ) and low error (RMSE = 1.725%), on the considered subset of 7493 samples (Figure 6e).



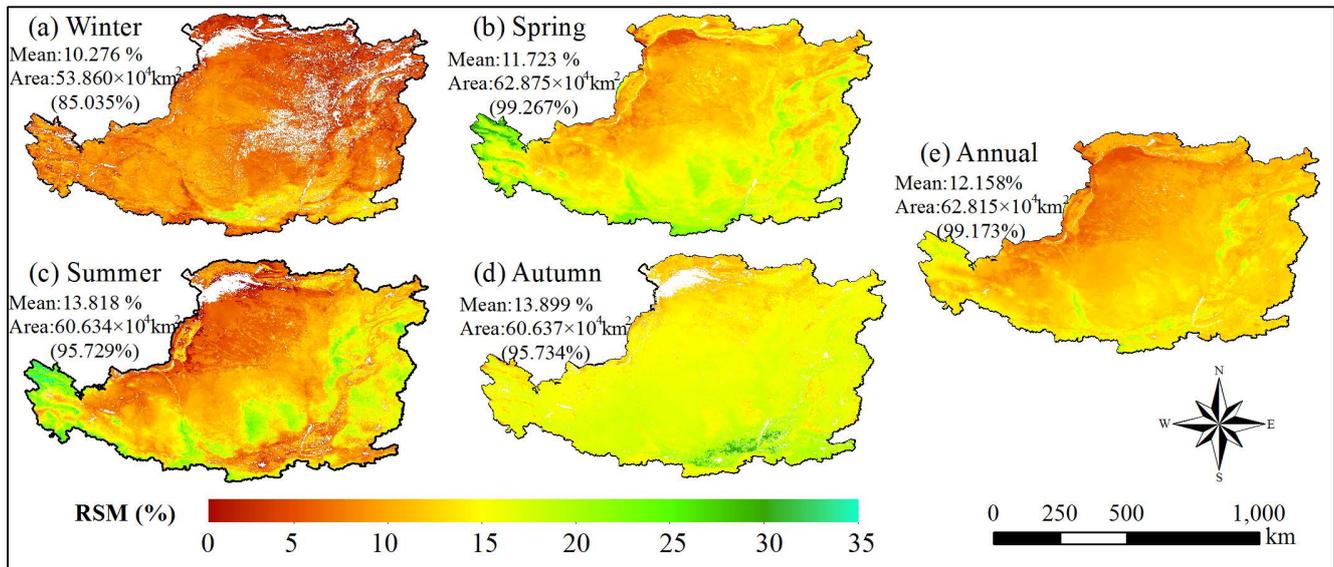
**Figure 6.** Model validation (modeled RSM by stepwise multilinear regression compared with reference RSM) on the subset selected at the seasonal and annual time scale. (a–d) Scatter plots of the reference and modeled RSM in four seasons of 2017; (e) scatter plot of the reference and modeled RSM in 2017. Scores (Pearson’s correlation coefficient ( $r$ ), Adjusted  $R^2$  (Adj.  $R^2$ ), standard deviation (STD), root mean square error (RMSE), and mean absolute error (MAE)) were computed using data included in the corresponding subplot boundary.  $N$  represents the number of available RSM samples for each month. The associated  $p$ -values (in the subplots) with the correlation coefficients are all  $< 0.001$ .

The model validation with the available sampled dataset at the monthly scale is presented in Figure A4. It should be noted that there were six months (December, January, February, April, May, and November) with Adj.  $R^2$  exceeding 0.800 in 2017 and relatively low errors (RMSE) of the models during the winter season (December, January, and February), ranging from 0.761% to 1.050%. Importantly, all intercepts and slopes of the linear fits to the modeled RSM were between 0 and 1, which indicates an overestimation of the model in the lower RSM region and underestimation of the model in the higher RSM region.

#### 4.3. Modeled Soil Moisture

Figure 7 depicts the spatial patterns of the seasonal and annual RSM over the CLP. Colors varying from red to blue indicate the change of RSM from lowest (i.e., driest soil) to highest (i.e., wettest soil). Overall, the spatiotemporal distribution of the modeled RSM via the SMLR model had the same pattern as the reference RSM. RSM decreased gradually from southeast to northwest and indicated an overall spatial distribution pattern of “wet in south and southeast, dry in northwest”. Such variation was also displayed in the annual PRE, RH, NL, and ET maps (Figure 2). The areas with low RSM (i.e., below 10%) were mainly clustered in the northwestern region of the CLP. These areas have a temperate

continental climate, with little rainfall, low RH, strong daylight, a high proportion of sand content, and low ET because of the low vegetation coverage. RSM was higher in the western region of the CLP (with the highest elevation) in spring and summer. The mean modeled RSM was highest (13.899%) in autumn and lowest (10.276%) in winter (Figure 7). The annual mean modeled RSM (12.158%) (Figure 7e), modeled via the SMLR model, was higher than that of the reference RSM (10.160%).



**Figure 7.** The spatial patterns of seasonal and annual RSM, modeled using SMLR over the CLP in 2017. (a–d) Spatial pattern of RSM in four seasons of 2017; (e) spatial pattern of RSM in 2017. The modeled RSM area and the mean RSM were computed for each season and the year 2017 in total. White color for each monthly RSM map means no value of RSM estimated via the SMLR method (the number in parentheses indicates the RSM areas in percentage against the whole area for each subfigure).

The monthly RSM maps from the multisource data modeled via the SMLR models are shown in Figure A5. Compared with the monthly reference RSM, a significant improvement was achieved in the spatial coverage of the modeled RSM via the SMLR model. The modeled RSM area of 10 months (except for January and February) was larger than that of the reference RSM in 2017. To be more specific, the area of modeled RSM increased more than two-fold over five months (April: 142.082%, May: 113.209%, July: 149.955%, August: 115.120%, and October: 243.008%) and the increased area of the modeled RSM ranged from  $1.020 \times 10^4$  km<sup>2</sup> in December to  $44.519 \times 10^4$  km<sup>2</sup> in October.

## 5. Discussion

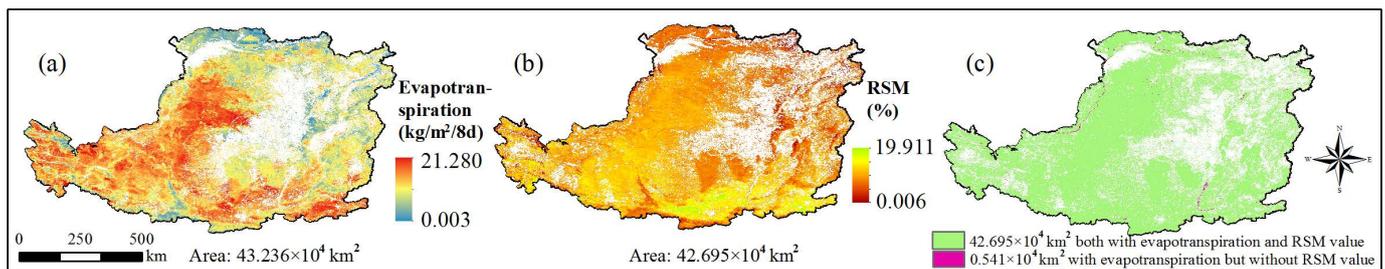
According to the results of the model fit, the highest Adj.  $R^2$  of 0.912, both in calibration and in validation, was found in winter. Accordingly, regression models in December, January, and February demonstrated much better performances (Adj.  $R^2$  ranged from 0.853 to 0.939 in validation) compared with other months. These results were better than those that were obtained by Lee et al. [57] who used MLR for South Korea (where the  $R^2$  ranged from 0.17 to 0.63). The linear regression equation had the lowest Adj.  $R^2$  of 0.091 in October (Adj.  $R^2$  of 0.073 in validation), which was mainly the result of the poor quality of the reference RSM with the lowest  $r$  of 0.47 against in situ RSM observations among months [32]. Similarly, the validation results for autumn (winter) were considerably lower (higher) than the others. The modeled results using SMLR were highly related to the quality of the input data, especially the reference RSM we previously published. The reference RSM in autumn had the greatest error and had the highest Pearson's  $r$  in winter among the four seasons. This may be considered the reason for the similar validation

results in the current study. To explore the reason or accurately estimate SM, a better model can be developed through future works. Thus, the performance of the SMLR model was found to be sensitive to the quality of the reference data.

For individual variables, the directions of regression coefficients in the regression equations varied between periods. Theoretically, the variables of precipitation and RSM should maintain a positive correlation when only precipitation is considered as an independent variable. However, in an arid area, RSM often reaches its highest value after a heavier rain; if several small rain events occur instead, RSM at a 20 cm depth does not increase and even keeps decreasing, as it is affected by other variables (e.g., temperature) in certain land covers [37]. In addition, many variables were selected in the regression models that weakened the effects of the variation of precipitation on RSM (even having a negative impact) compared with variables with strong positive influence. The interpretation of each variable in the models is, however, complex, and specific relationships between variables and RSM were not examined in this study but their influence would deserve further study. From the modeling, it is worth noting that variables such as ST1, ST3, and DIL, retained the same directions in the regression models throughout the period. This indicates that these variables demonstrate a stable correlation with RSM and maintain their impact even when other variables change [70].

Dummy variables like LC1 (11 months except for August) and ST5 (9 months) had high frequencies of application in the regression models among months. This suggests that after the effects of other variables were considered, these dummy variables could be higher or lower (indicating a positive or negative effect, respectively, according to the directions of the corresponding coefficient in the models) than the dependent variable of RSM. In particular, the land cover of croplands scored 0.632 points higher, at the annual regression model (Table 7) on the modeling RSM, than the reference groups (i.e., other land covers). In general, quantitative or categorical variables in regression models often had interaction effects with each other [89]. In the present study, the regression models did not allow for the possibility that interaction might occur among variables (no interaction term exists in the regression models). Thus, the regression coefficient for each variable could be individually interpreted as a statistically significant ( $p$ -value < 0.05) dependent variable (RSM). Several researchers reported that the number and position of the dummy variables affect the fitting degree and estimation accuracy of the resulting models [68,117,118]. Chen et al. found that the dummy variable model did not differ in regional biomass estimation ability [119].

The overall spatial pattern of the modeled RSM (with high RSM in the southern regions and low RSM in the northwestern areas of the CLP throughout the period) via SMLR was in good agreement with previous studies [11,32,120]. As for the mean RSM, the average of the mean modeled RSM of 12 months was 13.114%, which slightly exceeded that of the reference RSM (12.155%) [32]. Regarding the overestimation of the dry area and the underestimation of the wet area, this might be partly attributed to the overestimation or underestimation of certain selected variables. The present study did not consider such parameter uncertainty, which was considered as a limitation of this study. In addition, the area of modeled RSM in January and February was smaller (larger for other periods) than that of the reference RSM. Each of the selected variables in the regression models (i.e., the value and distribution) were analyzed, which showed that this was associated with the selected variable of ET. The smaller areas of modeled RSM in January and February were caused by the ET maps that were still incomplete at that time. Taking January as an example, the areas of ET and modeled RSM were  $43.236 \times 10^4 \text{ km}^2$  and  $42.695 \times 10^4 \text{ km}^2$ , respectively (Figure 8). The region with modeled RSM, as shown in Figure 8c, also had ET value, and the area of  $0.541 \times 10^4 \text{ km}^2$  with ET but without an RSM value should be associated with other unavailable variables (e.g., DL) at that time. Pixels with a DL below  $0^\circ \text{C}$  were removed to avoid freezing temperatures in winter [104]. Therefore, a limitation of the SMLR model, to some extent, might be the fact that the availability of independent variables directly affects the coverage of the modeled RSM.



**Figure 8.** Comparison between the area of the evapotranspiration and modeled RSM for January over the CLP. (a) The spatial pattern of evapotranspiration with an area of  $43.236 \times 10^4 \text{ km}^2$  in January; (b) spatial pattern of modeled RSM with an area of  $42.695 \times 10^4 \text{ km}^2$  in January; (c) different spatial distributions of evapotranspiration and modeled RSM. The area in green color ( $42.695 \times 10^4 \text{ km}^2$ ) represents the area of modeled RSM and the pixels ( $0.541 \times 10^4 \text{ km}^2$ ) with the evapotranspiration value but without modeled RSM are illustrated in the magenta color.

According to the outcomes of the validation, for six months (December, January, February, April, May, and November), the Adj.  $R^2$  exceeded 0.800 in 2017. These results proved the effectiveness of the SMLR method throughout the year in the study area [67,69]. This suggests that the SMLR method is a promising approach to estimate RSM. To retrieve RSM in each area and period using the SMLR method, the selected variables and coefficients of these variables only need to be updated. From a practical point of view, this is a significant finding, as it supports the use of multisource data to complement and/or replace coarse resolution satellite imagery in the simulation of RSM.

However, the methods used in the present study induce uncertainties. Only 17 features were used because of data limitations and hydrological parameters (e.g., runoff and irrigation activities), and other soil properties (e.g., soil porosity, bulk density, and soil organic matter) were ignored since these were difficult to obtain. There is no guarantee that the modeling outcomes will become better if more variables are collected. Certain variables could be removed when a new variable is added because of the issue of multicollinearity and the significance level. For regions where the availability of input variables is restricted, the simplest model, which includes soil texture and organic carbon, might be an alternative for estimating the water availability in the soil [58]. In addition, because the performance of the SMLR model was sensitive to the quality of the reference data as well as input data, the study was limited by the accuracy of the used reference RSM. Moreover, the good quality of each pixel for the inputs would improve the accuracy of the modeled RSM. The MODIS inputs (except for the land cover) were composite data over each 8-day period and could be affected by clouds or atmospheric interference. Besides ensuring each pixel value within the valid range, quality control procedures should be performed for each dataset using the quality flags in future studies [121,122].

## 6. Conclusions

Based on reference RSM (relative soil moisture (obtained in the previous study)), 34 candidate variables (12 quantitative variables and 22 dummy variables) from multi-source data, including Moderate Resolution Imaging Spectroradiometer (MODIS) and topographic data, soil properties data, and meteorological data, were processed and modeled via stepwise multilinear regression (SMLR). After the accuracy assessment, monthly, seasonal, and annual spatial patterns of the modeled RSM were mapped and evaluated over the Chinese Loess Plateau (CLP) in 2017. The key findings and main conclusions are summarized in the following:

- SMLR could model RSM with the desired accuracy ( $r = 0.969$ ,  $RMSE = 0.761\%$ ,  $MAE = 0.576\%$  in December) at a 500 m resolution over the CLP. Moreover, the use of multisource data to complement and/or replace coarse-resolution satellite data in RSM modeling could be considered.

- The variables of elevation (0–500 m and 2000–2500 m), precipitation, soil texture of the loam, and nighttime land surface temperature could continuously be used in the regression models for all seasons in 2017.
- Including dummy variables improved the model fit, both in calibration and validation, because the Adj.  $R^2$  was higher with dummy variables than without.
- The SMLR-modeled RSM for almost all periods except for January and February (because of unavailable ET data at that time) achieved better spatial coverage than that of the reference RSM. Thus, the availability of selected variables directly affects the coverage of the modeled RSM.

The SMLR-modeled RSM successfully characterized the spatiotemporal variability of RSM and agreed well with the reference RSM. The modeled RSM maps generated in this study are feasible for further study and can be regarded as in situ SM data. Further studies and validation of the presented SMLR models are advisable, which can be achieved by extending the investigation to other datasets, test areas, and data periods.

**Author Contributions:** Conceptualization, Lina Yuan and Long Li; methodology, Lina Yuan; software, Weiqiang Liu and Sai Hu; validation, Lina Yuan, and Long Li; resources, Ting Zhang and Longhua Yang; writing—original draft preparation, Lina Yuan, and Ting Zhang; writing—review and editing, Long Li and Longqian Chen; visualization, Lina Yuan, and Sai Hu; supervision, Longqian Chen and Long Li. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by the Fundamental Research Funds for the Central Universities (Grant No.: 2018ZDPY07).

**Institutional Review Board Statement:** Not applicable for this study.

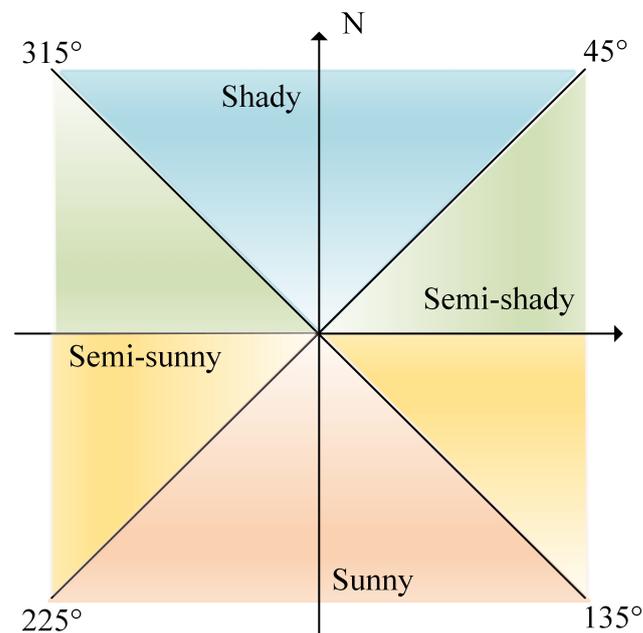
**Informed Consent Statement:** Not applicable for this study.

**Data Availability Statement:** The input data used for this study can be accessed freely from on-line sources.

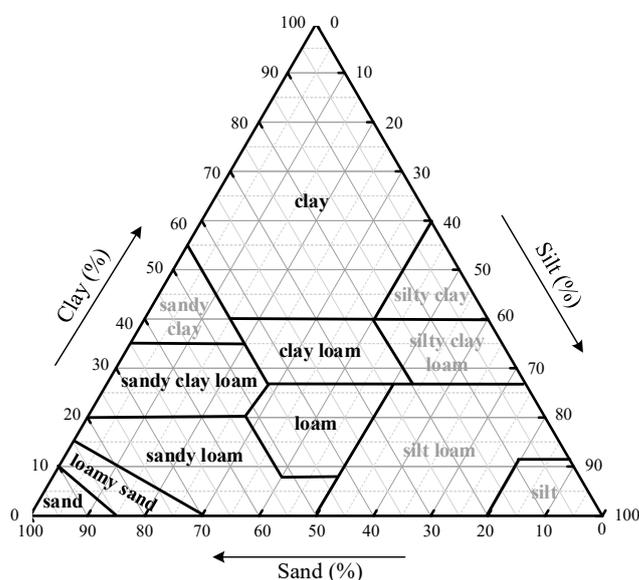
**Acknowledgments:** The authors would like to thank the China Meteorological Data Service Center (CMDSC, <http://data.cma.cn/en>, accessed on 11 January 2019), the NASA Land Processes Distributed Active Archive Center (LP DAAC, <https://lpdaac.usgs.gov/>, accessed on 21 February 2019), the Data Center for Resource and Environmental Sciences, Chinese Academy of Sciences (RESDC, <http://www.resdc.cn/>, accessed on 20 August 2020), the U.S. Department of Agriculture (USDA, <https://www.usda.gov/>, accessed on 21 September 2020), and the Geospatial Data Cloud Platform (<http://www.gscloud.cn/>, accessed on 11 October 2020) for providing the data required for the study. Furthermore, we appreciate the editors and reviewers for their constructive comments and suggestions.

**Conflicts of Interest:** The authors declare no conflict of interest.

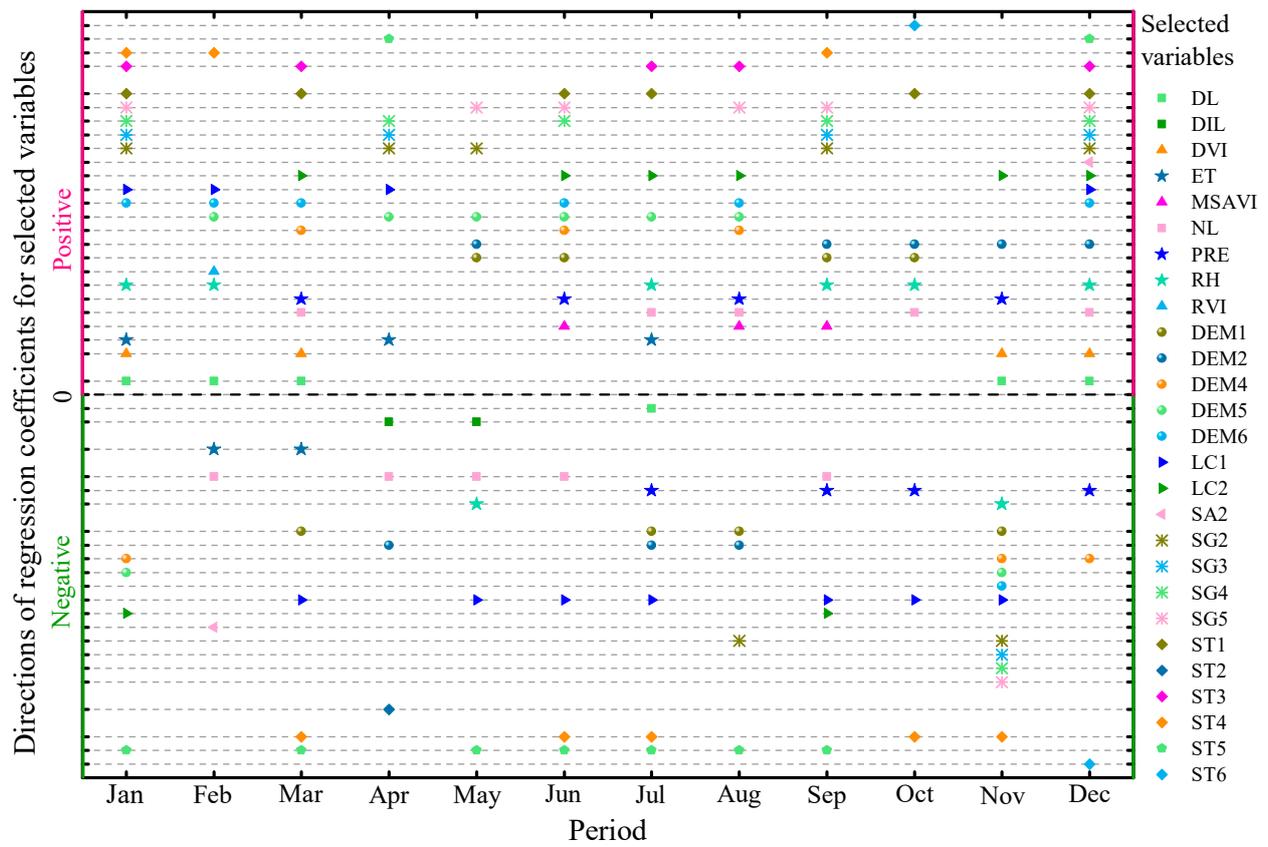
## Appendix A



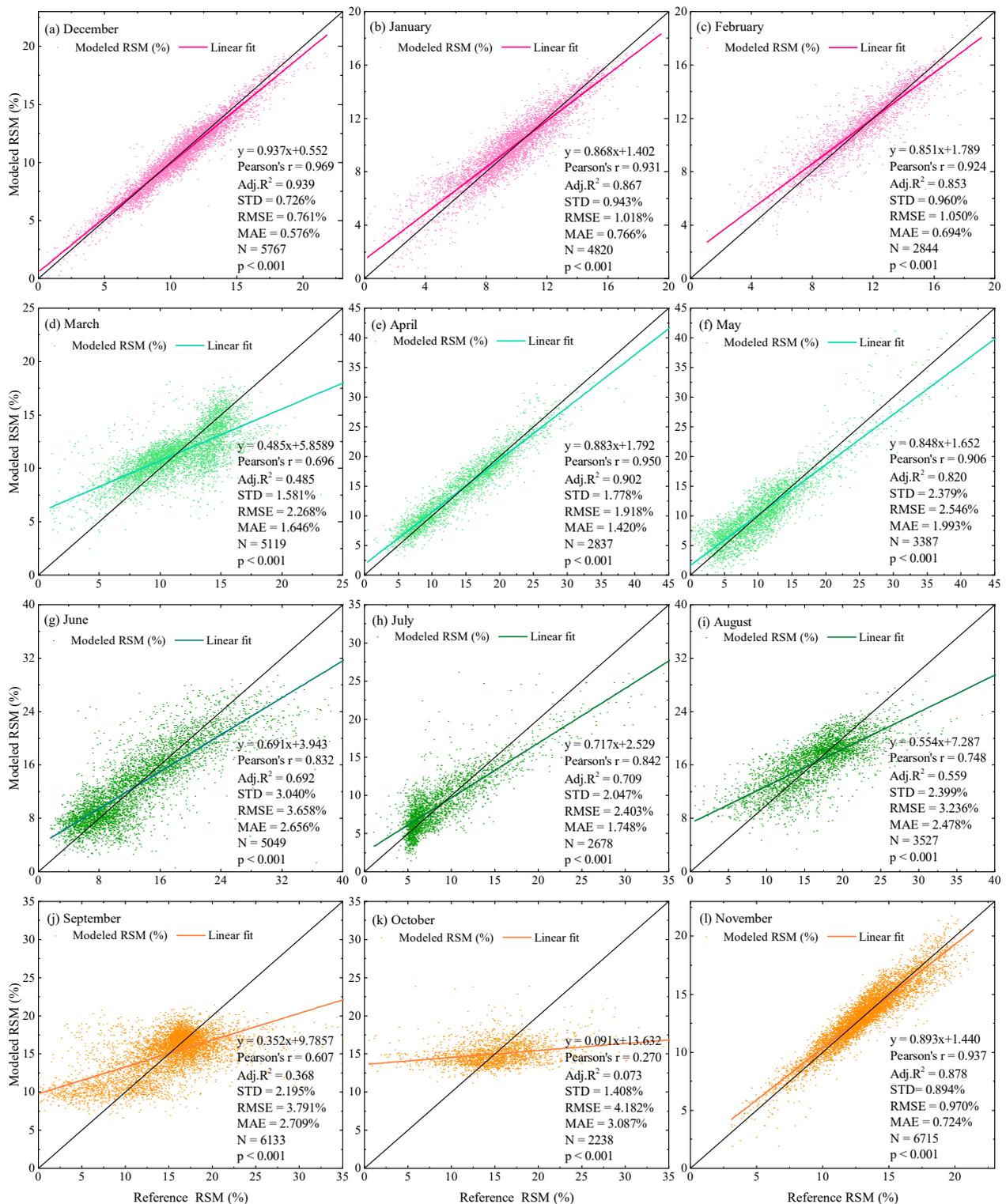
**Figure A1.** Schematic diagram of the four directions of the slope aspects (adapted from [123]). According to the intensity of solar radiation, the slope aspects ( $0^{\circ}$ – $360^{\circ}$ ) in the Northern Hemisphere were divided into shady ( $0^{\circ}$ – $45^{\circ}$  and  $315^{\circ}$ – $360^{\circ}$ ), semi-shady ( $45^{\circ}$ – $90^{\circ}$  and  $270^{\circ}$ – $315^{\circ}$ ), sunny ( $135^{\circ}$ – $225^{\circ}$ ), and semi-sunny ( $90^{\circ}$ – $135^{\circ}$  and  $225^{\circ}$ – $270^{\circ}$ ) directions.



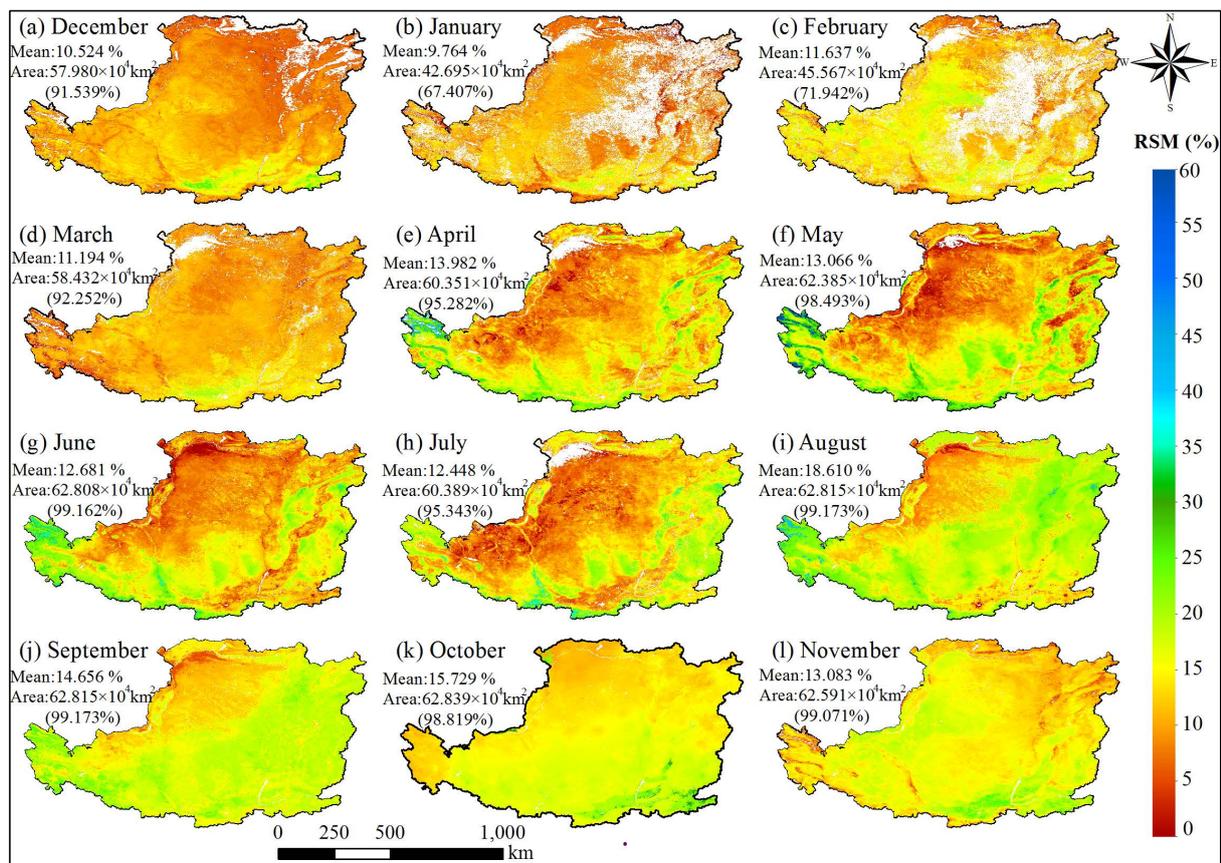
**Figure A2.** Soil texture triangular classification diagram used to classify soil properties (adapted from a figure provided by the U.S. Department of Agriculture (USDA) at <https://www.usda.gov/>, accessed on 21 September 2020). The soil properties are determined via the composition of the soil (i.e., the percentages of sand, silt, and clay). The summation of the percentages of sand, silt, and clay equals 100%. The corners of the triangle indicate 100% of each composition, and 12 textural classes of soil are noted within the triangle using thick lines as separations between classes. Soil textures over the CLP are formatted in bold black (seven textural classes), and soil textures shown in bold grey do not appear in the study area.



**Figure A3.** Selected variables for modeling at the monthly scale. Labels above and below 0 represent positive and negative directions of regression coefficients for selected variables in the regression models, respectively.



**Figure A4.** Model validation (modeled RSM by SMLR compared with reference RSM) on the subset selected at the monthly time scale. Scores (Pearson's correlation coefficient ( $r$ ), Adjusted  $R^2$  (Adj.  $R^2$ ), standard deviation (STD), root mean square error (RMSE), and mean absolute error (MAE)) were computed using data included in the corresponding subplot boundary.  $N$  represents the number of available RSM samples for each month. The associated  $p$ -values (in the subplots) with the correlation coefficients are all < 0.001. The colors of modeled RSM and linear fit are pink, light green, dark green, and orange representing winter, spring, summer, and autumn, respectively.



**Figure A5.** The spatial pattern of monthly RSM modeled using SMLR over the CLP in 2017. Modeled RSM area and mean RSM were computed for each month. White color for each monthly RSM map means no value of RSM estimated via the SMLR method (the number in parentheses indicates the RSM areas in percentage against the whole area for each subfigure).

## Appendix B

**Table A1.** Information of the reference relative soil moisture (RSM) regarding the accuracy assessment against in situ RSM observation (station-based) as well as the mean RSM and the area of the reference RSM for each period.

Period	Validation Results				Reference RSM	
	r	Adj. R <sup>2</sup>	MAE (%)	RMSE (%)	Mean RSM (%)	Area (10 <sup>4</sup> km <sup>2</sup> )
Annual	0.73	0.52	3.00	3.75	10.16	63.24
Winter	0.67	0.44	3.22	3.74	9.08	60.90
Spring	0.53	0.27	3.14	3.98	11.68	58.00
Summer	0.58	0.34	3.25	3.86	13.82	54.57
Autumn	0.67	0.44	3.64	4.41	13.91	61.99
January	0.68	0.45	3.24	4.06	9.31	56.18
February	0.66	0.42	3.23	3.84	8.64	62.39
March	0.57	0.32	3.02	3.91	11.42	42.92
April	0.64	0.41	3.09	3.77	14.18	24.93
May	0.65	0.42	3.27	3.89	10.50	29.26
June	0.59	0.35	3.81	4.54	12.83	43.44
July	0.54	0.28	3.43	4.29	8.85	24.16
August	0.61	0.37	3.62	4.48	16.42	29.20
September	0.57	0.32	3.58	4.57	14.91	51.91
October	0.47	0.21	4.97	6.10	15.05	18.32
November	0.50	0.25	3.80	4.66	13.03	61.01
December	0.64	0.40	3.26	3.99	10.72	56.96

**Table A2.** Formulas used to calculate vegetation indexes [98,124,125].

Vegetation Index	Formula
Enhanced vegetation index (EVI)	$EVI = \frac{2.5 \times (\rho_{NIR} - \rho_{Red})}{\rho_{NIR} + 6\rho_{Red} - 7.5\rho_{Blue} + 1}$
Difference vegetation index (DVI)	$DVI = \rho_{NIR} - \rho_{Red}$
Ratio vegetation index (RVI)	$RVI = \frac{\rho_{NIR}}{\rho_{Red}}$
Normalized difference vegetation index (NDVI)	$NDVI = \frac{\rho_{NIR} - \rho_{Red}}{\rho_{NIR} + \rho_{Red}}$
Enhanced vegetation index 2 (EVI2)	$EVI2 = \frac{\rho_{NIR} - \rho_{Red}}{\rho_{NIR} + 2.4\rho_{Red} + 1}$
Modified soil-adjusted vegetation index (MSAVI)	$MSAVI = \frac{2\rho_{NIR} + 1 - \sqrt{(2\rho_{NIR} + 1)^2 - 8(\rho_{NIR} - \rho_{Red})}}{2}$

**Table A3.** Mathematical expressions of the goodness of validation [11].

Abbreviation	Formula <sup>1</sup>
RMSE	$\sqrt{\frac{\sum_{i=1}^n (O_i - P_i)^2}{n}}$
r	$\frac{\sum_{i=1}^n (O_i - \bar{O})(P_i - \bar{P})}{\sqrt{\sum_{i=1}^n (O_i - \bar{O})^2} \sqrt{\sum_{i=1}^n (P_i - \bar{P})^2}}$
MAE	$\frac{\sum_{i=1}^n  O_i - P_i }{n}$
R <sup>2</sup>	$\left( \frac{\sum_{i=1}^n (O_i - \bar{O})(P_i - \bar{P})}{\sqrt{\sum_{i=1}^n (O_i - \bar{O})^2} \sqrt{\sum_{i=1}^n (P_i - \bar{P})^2}} \right)^2$
STD	$\sqrt{\frac{\sum_{i=1}^n (P_i - \bar{P})^2}{n}}$

<sup>1</sup>P and O represent the model- and reference-based RSM, respectively. N represents the total number of pairs for assessment, and i represents the i<sup>th</sup> sample.  $\bar{P}$  and  $\bar{O}$  represent the mean value of model- and reference-based RSM, respectively.

**Table A4.** Stepwise multilinear regression analysis of monthly RSM.

Period	Regression Model	Adj. R <sup>2</sup>	RMSE	Max VIF
January	RSM = −243.006 + 0.900*DL + 15.761*DVI + 0.150*ET + 0.021*RH − 0.350*DEM4 − 0.153*DEM5 + 0.508*DEM6 + 0.285*LC1 − 0.630*LC2 + 0.275*SG2 + 0.448*SG3 + 0.423*SG4 + 0.513*SG5 + 0.170*ST1 + 0.816*ST3 + 0.529*ST4 − 0.221*ST5	0.864	1.016	1.884
February	RSM = −165.850 + 0.688*DL − 0.059*ET − 0.072*NL + 0.031*RH + 0.502*RVI + 0.341*DEM5 + 0.442*DEM6 + 0.214*LC1 − 0.104*SA2 + 0.338*ST4	0.849	1.034	1.733
March	RSM = −139.430 + 0.431*DL + 20.424*DVI − 0.144*ET + 0.096*NL + 0.033*PRE − 0.375*DEM1 + 0.426*DEM4 + 1.114*DEM6 − 0.433*LC1 + 3.674*LC2 + 0.266*ST1 + 2.537*ST3 − 1.147*ST4 − 1.191*ST5	0.467	2.317	2.382
April	RSM = 257.901 − 1.070*DIL + 0.183*ET − 0.803*NL − 0.152*DEM2 + 0.179*DEM5 + 0.620*LC1 + 0.414*SG2 + 0.769*SG3 + 0.673*SG4 − 1.131*ST2 + 0.577*ST5	0.899	1.892	2.138
May	RSM = 361.380 − 1.130*DIL − 1.132*NL − 0.078*RH + 1.072*DEM1 + 0.502*DEM2 + 0.420*DEM5 − 0.277*LC1 + 0.276*SG2 + 0.512*SG5 − 1.039*ST5	0.823	2.545	1.938
June	RSM = 222.274 + 19.787*MSAVI − 0.764*NL + 0.044*PRE + 0.813*DEM1 + 0.539*DEM4 + 0.863*DEM5 + 1.460*DEM6 − 0.437*LC1 + 4.148*LC2 + 0.387*SG4 + 0.989*SG5 + 0.396*ST1 − 3.075*ST4 − 1.648*ST5	0.708	3.528	1.993
July	RSM = 160.019 − 0.648*DL + 0.141*ET + 0.154*NL − 0.016*PRE + 0.116*RH − 1.260*DEM1 − 0.585*DEM2 + 0.393*DEM5 − 0.622*LC1 + 2.584*LC2 + 0.233*ST1 + 5.579*ST3 − 0.853*ST4 − 0.368*ST5	0.694	2.434	2.604
August	RSM = 167.362 + 19.965*MSAVI − 0.548*NL + 0.018*PRE − 1.666*DEM1 − 0.865*DEM2 + 1.372*DEM4 + 1.845*DEM5 + 1.297*DEM6 + 2.000*LC2 − 0.314*SG2 + 0.618*SG5 + 9.771*ST3 − 1.965*ST5	0.560	3.197	2.856

Table A4. Cont.

Period	Regression Model	Adj. R <sup>2</sup>	RMSE	Max VIF
September	RSM = 83.292 + 13.956*MSAVI − 0.278*NL − 0.006*PRE + 0.079*RH + 3.487*DEM1 + 1.625*DEM2 − 0.575*LC1 − 2.252*LC2 + 0.866*SG2 + 0.669*SG3 + 0.789*SG4 + 0.470*SG5 + 1.810*ST4 − 0.866*ST5	0.334	3.926	2.841
October	RSM = −31.338 + 0.124*NL − 0.023*PRE + 0.191*RH + 3.825*DEM1 + 0.923*DEM2 − 0.817*LC1 + 0.505*ST1 − 0.913*ST4 + 9.032*ST6	0.091	4.182	1.592
November	RSM = −255.132 + 0.948*DL + 8.246*DVI + 0.021*PRE − 0.007*RH − 0.429*DEM1 + 0.100*DEM2 − 0.342*DEM4 − 0.423*DEM5 − 0.203*DEM6 − 0.196*LC1 + 0.238*LC2 − 0.177*SG2 − 0.137*SG3 − 0.166*SG4 − 0.144*SG5 − 0.421*ST4	0.880	0.938	2.345
December	RSM = −274.446 + 0.999*DL + 21.603*DVI + 0.021*NL − 0.036*PRE + 0.020*RH + 0.125*DEM2 − 0.163*DEM4 + 0.137*DEM6 + 0.069*LC1 + 0.617*LC2 + 0.052*SA2 + 0.122*SG2 + 0.152*SG3 + 0.203*SG4 + 0.121*SG5 + 0.065*ST1 + 0.875*ST3 + 0.219*ST5 − 0.846*ST6	0.938	0.753	2.293

## References

- Albertson, J.D.; Kiely, G. On the structure of soil moisture time series in the context of land surface models. *J. Hydrol.* **2001**, *243*, 101–119. [\[CrossRef\]](#)
- Spennemann, P.C.; Salvia, M.; Ruscica, R.C.; Sörensson, A.A.; Grings, F.; Karszenbaum, H. Land-atmosphere interaction patterns in southeastern South America using satellite products and climate models. *Int. J. Appl. Earth Obs. Geoinf.* **2018**, *64*, 96–103. [\[CrossRef\]](#)
- Tayfur, G.; Zucco, G.; Brocca, L.; Moramarco, T. Coupling soil moisture and precipitation observations for predicting hourly runoff at small catchment scale. *J. Hydrol.* **2014**, *510*, 363–371. [\[CrossRef\]](#)
- Zhang, B.; Aghakouchak, A.; Yang, Y.; Wei, J.; Wang, G. A water-energy balance approach for multi-category drought assessment across globally diverse hydrological basins. *Agric. For. Meteorol.* **2019**, *264*, 247–265. [\[CrossRef\]](#)
- Dorigo, W.; de Jeu, R. Satellite soil moisture for advancing our understanding of earth system processes and climate change. *Int. J. Appl. Earth Obs. Geoinf.* **2016**, *48*, 1–4. [\[CrossRef\]](#)
- Kumar, S.V.; Dirmeyer, P.A.; Peters-Lidard, C.D.; Bindlish, R.; Bolten, J. Information theoretic evaluation of satellite soil moisture retrievals. *Remote Sens. Environ.* **2018**, *204*, 392–400. [\[CrossRef\]](#)
- Liu, D.; Mishra, A.K.; Yu, Z.; Yang, C.; Konapala, G.; Vu, T. Performance of SMAP, AMSR-E and LAI for weekly agricultural drought forecasting over continental United States. *J. Hydrol.* **2017**, *553*, 88–104. [\[CrossRef\]](#)
- Kim, S.; Zhang, R.; Pham, H.; Sharma, A. A review of satellite-derived soil moisture and its usage for flood estimation. *Remote Sens. Earth Syst. Sci.* **2019**, *2*, 225–246. [\[CrossRef\]](#)
- Singh, A.K.; Bhardwaj, A.K.; Verma, C.L.; Mishra, V.K. Soil moisture sensing techniques for scheduling irrigation. *J. Soil Salin. Water Qual.* **2019**, *11*, 68–76.
- Zaussinger, F.; Dorigo, W.; Gruber, A.; Tarpanelli, A.; Filippucci, P.; Brocca, L. Estimating irrigation water use over the contiguous United States by combining satellite and reanalysis soil moisture data. *Hydrol. Earth Syst. Sci.* **2019**, *23*, 897–923. [\[CrossRef\]](#)
- Yuan, L.; Li, L.; Zhang, T.; Chen, L.; Zhao, J.; Hu, S.; Cheng, L.; Liu, W. Soil moisture estimation for the Chinese Loess Plateau using MODIS-derived ATI and TVDI. *Remote Sens.* **2020**, *12*, 3040. [\[CrossRef\]](#)
- Fang, B.; Lakshmi, V. Soil moisture at watershed scale: Remote sensing techniques. *J. Hydrol.* **2014**, *516*, 258–272. [\[CrossRef\]](#)
- Ma, H.; Zhang, L.; Sun, L.; Liu, Q. Farmland soil moisture inversion by synergizing optical and microwave remote sensing data. *J. Remote Sens.* **2014**, *18*, 673–685. [\[CrossRef\]](#)
- Karthikeyan, L.; Pan, M.; Wanders, N.; Kumar, D.N.; Wood, E.F. Four decades of microwave satellite soil moisture observations: Part 2. Product validation and inter-satellite comparisons. *Adv. Water Resour.* **2017**, *109*, 236–252. [\[CrossRef\]](#)
- Djamai, N.; Magagi, R.; Goïta, K.; Merlin, O.; Kerr, Y.; Roy, A. A combination of DISPATCH downscaling algorithm with CLASS land surface scheme for soil moisture estimation at fine scale during cloudy days. *Remote Sens. Environ.* **2016**, *184*, 1–14. [\[CrossRef\]](#)
- Piles, M.; Ballabrera-Poy, J.; Muñoz-Sabater, J. Dominant features of global surface soil moisture variability observed by the SMOS satellite. *Remote Sens.* **2019**, *11*, 95. [\[CrossRef\]](#)
- Doubková, M.; van Dijk, A.I.J.M.; Sabel, D.; Wagner, W.; Blöschl, G. Evaluation of the predicted error of the soil moisture retrieval from C-band SAR by comparison against modelled soil moisture estimates over Australia. *Remote Sens. Environ.* **2012**, *120*, 188–196. [\[CrossRef\]](#)
- Han, J.; Mao, K.; Xu, T.; Guo, J.; Zuo, Z.; Gao, C. A soil moisture estimation framework based on the CART algorithm and its application in China. *J. Hydrol.* **2018**, *563*, 65–75. [\[CrossRef\]](#)
- Dumedah, G.; Walker, J.P. Intercomparison of the JULES and CABLE land surface models through assimilation of remotely sensed soil moisture in southeast Australia. *Adv. Water Resour.* **2014**, *74*, 231–244. [\[CrossRef\]](#)
- Moon, H.; Choi, M. Dryness indices based on remotely sensed vegetation and land surface temperature for evaluating the soil moisture status in cropland-forest-dominant watersheds. *Terr. Atmos. Ocean. Sci.* **2015**, *26*, 599–611. [\[CrossRef\]](#)

21. Malbêteau, Y.; Merlin, O.; Molero, B.; Rüdiger, C.; Bacon, S. DisPATCH as a tool to evaluate coarse-scale remotely sensed soil moisture using localized in situ measurements: Application to SMOS and AMSR-E data in Southeastern Australia. *Int. J. Appl. Earth Obs. Geoinf.* **2016**, *45*, 221–234. [[CrossRef](#)]
22. Djamaï, N.; Magagi, R.; Goita, K.; Merlin, O.; Kerr, Y.; Walker, A. Disaggregation of SMOS soil moisture over the Canadian Prairies. *Remote Sens. Environ.* **2015**, *170*, 255–268. [[CrossRef](#)]
23. Tagesson, T.; Horion, S.; Nieto, H.; Zaldo Fornies, V.; Mendiguren González, G.; Bulgin, C.E.; Ghent, D.; Fensholt, R. Disaggregation of SMOS soil moisture over West Africa using the Temperature and Vegetation Dryness Index based on SEVIRI land surface parameters. *Remote Sens. Environ.* **2018**, *206*, 424–441. [[CrossRef](#)]
24. Merlin, O.; Malbêteau, Y.; Notfi, Y.; Bacon, S.; Er-raki, S. Performance metrics for soil moisture downscaling methods: Application to DISPATCH data in Central Morocco. *Remote Sens.* **2015**, *7*, 3783–3807. [[CrossRef](#)]
25. Zhang, D.; Zhou, G. Estimation of soil moisture from optical and thermal remote sensing: A review. *Sensors* **2016**, *16*, 1308. [[CrossRef](#)]
26. Liu, Y.; Qian, J.; Yue, H. Combined Sentinel-1A with Sentinel-2A to estimate soil moisture in farmland. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 1292–1310. [[CrossRef](#)]
27. Koley, S.; Jeganathan, C. Estimation and evaluation of high spatial resolution surface soil moisture using multi-sensor multi-resolution approach. *Geoderma* **2020**, *378*, 114618. [[CrossRef](#)]
28. Palombo, A.; Pascucci, S.; Loperte, A.; Lettino, A.; Castaldi, F.; Muolo, M.R.; Santini, F. Soil moisture retrieval by integrating TASI-600 airborne thermal data, WorldView 2 satellite data and field measurements: Petacciato case study. *Sensors* **2019**, *19*, 1515. [[CrossRef](#)] [[PubMed](#)]
29. Wang, H.; He, N.; Zhao, R.; Ma, X. Soil water content monitoring using joint application of PDI and TVDI drought indices. *Remote Sens. Lett.* **2020**, *11*, 455–464. [[CrossRef](#)]
30. Lu, X.J.; Zhou, B.; Yan, H.B.; Luo, L.; Huang, Y.H.; Wu, C.L. Remote sensing retrieval of soil moisture in Guangxi based on ATI and TVDI models. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. ISPRS Arch.* **2020**, *42*, 895–902. [[CrossRef](#)]
31. Lu, L.; Luo, G.P.; Wang, J.Y. Development of an ATI-NDVI method for estimation of soil moisture from MODIS data. *Int. J. Remote Sens.* **2014**, *35*, 3797–3815. [[CrossRef](#)]
32. Yuan, L.; Li, L.; Zhang, T.; Chen, L.; Zhao, J.; Liu, W.; Cheng, L.; Hu, S.; Yang, L.; Wen, M. Improving soil moisture estimation by identification of NDVI thresholds optimization: An application to the Chinese Loess Plateau. *Remote Sens.* **2021**, *13*, 589. [[CrossRef](#)]
33. Sabaghy, S.; Walker, J.P.; Renzullo, L.J.; Jackson, T.J. Spatially enhanced passive microwave derived soil moisture: Capabilities and opportunities. *Remote Sens. Environ.* **2018**, *209*, 551–580. [[CrossRef](#)]
34. Rahimi-Ajdadi, F.; Abbaspour-Gilandeh, Y.; Mollazade, K.; Hasanzadeh, R.P.R. Development of a novel machine vision procedure for rapid and non-contact measurement of soil moisture content. *Meas. J. Int. Meas. Confed.* **2018**, *121*, 179–189. [[CrossRef](#)]
35. Sandells, M.J.; Davenport, I.J.; Gurney, R.J. Passive L-band microwave soil moisture retrieval error arising from topography in otherwise uniform scenes. *Adv. Water Resour.* **2008**, *31*, 1433–1443. [[CrossRef](#)]
36. Fu, B.; Wang, J.; Chen, L.; Qiu, Y. The effects of land use on soil moisture variation in the Danangou catchment of the Loess Plateau, China. *Catena* **2003**, *54*, 197–213. [[CrossRef](#)]
37. Niu, C.Y.; Musa, A.; Liu, Y. Analysis of soil moisture condition under different land uses in the arid region of Horqin sandy land, northern China. *Solid Earth* **2015**, *6*, 1157–1167. [[CrossRef](#)]
38. Jiao, L.; Lu, N.; Fu, B.; Wang, J.; Li, Z.; Fang, W.; Liu, J.; Wang, C.; Zhang, L. Evapotranspiration partitioning and its implications for plant water use strategy: Evidence from a black locust plantation in the semi-arid Loess Plateau, China. *For. Ecol. Manag.* **2018**, *424*, 428–438. [[CrossRef](#)]
39. Maheu, A.; Ancil, F.; Gaborit, É.; Fortin, V.; Nadeau, D.F.; Therrien, R. A field evaluation of soil moisture modelling with the Soil, Vegetation, and Snow (SVS) land surface model using evapotranspiration observations as forcing data. *J. Hydrol.* **2018**, *558*, 532–545. [[CrossRef](#)]
40. Xu, Q.; Zhou, B. Retrieving soil water contents from soil temperature measurements by using linear regression. *Adv. Atmos. Sci.* **2003**, *20*, 849–858. [[CrossRef](#)]
41. Wang, Y.; Shao, M.; Zhu, Y.; Sun, H.; Fang, L. A new index to quantify dried soil layers in water-limited ecosystems: A case study on the Chinese Loess Plateau. *Geoderma* **2018**, *322*, 1–11. [[CrossRef](#)]
42. Li, X.; Xu, X.; Liu, W.; He, L.; Zhang, R.; Xu, C.; Wang, K. Similarity of the temporal pattern of soil moisture across soil profile in karst catchments of southwestern China. *J. Hydrol.* **2017**, *555*, 659–669. [[CrossRef](#)]
43. Abowarda, A.S.; Bai, L.; Zhang, C.; Long, D.; Li, X.; Huang, Q.; Sun, Z. Generating surface soil moisture at 30 m spatial resolution using both data fusion and machine learning toward better water resources management at the field scale. *Remote Sens. Environ.* **2021**, *255*, 112301. [[CrossRef](#)]
44. Khaki, M.; Zerihun, A.; Awange, J.L.; Dewan, A. Integrating satellite soil-moisture estimates and hydrological model products over Australia. *Aust. J. Earth Sci.* **2020**, *67*, 265–277. [[CrossRef](#)]
45. Liu, D.; Mishra, A.K. Performance of AMSR\_E soil moisture data assimilation in CLM4.5 model for monitoring hydrologic fluxes at global scale. *J. Hydrol.* **2017**, *547*, 67–79. [[CrossRef](#)]
46. Zhao, L.; Yang, K.; Qin, J.; Chen, Y.; Tang, W.; Lu, H.; Yang, Z.L. The scale-dependence of SMOS soil moisture accuracy and its improvement through land data assimilation in the central Tibetan Plateau. *Remote Sens. Environ.* **2014**, *152*, 345–355. [[CrossRef](#)]

47. Li, F.; Crow, W.T.; Kustas, W.P. Towards the estimation root-zone soil moisture via the simultaneous assimilation of thermal and microwave soil moisture retrievals. *Adv. Water Resour.* **2010**, *33*, 201–214. [[CrossRef](#)]
48. Bayat, A.T.; Schonbrodt-Stitt, S.; Nasta, P.; Ahmadian, N.; Conrad, C.; Bogena, H.R.; Vereecken, H.; Jakobi, J.; Baatz, R.; Romano, N. Mapping near-surface soil moisture in a Mediterranean agroforestry ecosystem using Cosmic-Ray Neutron Probe and Sentinel-1 Data. In Proceedings of the 2020 IEEE International Workshop on Metrology for Agriculture and Forestry (MetroAgriFor), Trento, Italy, 4–6 November 2020; Volume 1, pp. 201–206. [[CrossRef](#)]
49. Liu, Y.; Yao, L.; Jing, W.; Di, L.; Yang, J.; Li, Y. Comparison of two satellite-based soil moisture reconstruction algorithms: A case study in the state of Oklahoma, USA. *J. Hydrol.* **2020**, *590*, 125406. [[CrossRef](#)]
50. Dumedah, G.; Walker, J.P.; Chik, L. Assessing artificial neural networks and statistical methods for infilling missing soil moisture records. *J. Hydrol.* **2014**, *515*, 330–344. [[CrossRef](#)]
51. Santi, E.; Paloscia, S.; Pettinato, S.; Brocca, L.; Ciabatta, L.; Entekhabi, D. On the synergy of SMAP, AMSR2 and SENTINEL-1 for retrieving soil moisture. *Int. J. Appl. Earth Obs. Geoinf.* **2018**, *65*, 114–123. [[CrossRef](#)]
52. Elshorbagy, A.; Parasuraman, K. On the relevance of using artificial neural networks for estimating soil moisture content. *J. Hydrol.* **2008**, *362*, 1–18. [[CrossRef](#)]
53. Sharma, S.; Ochsner, T.E.; Twidwell, D.; Carlson, J.D.; Krueger, E.S.; Engle, D.M.; Fuhlendorf, S.D. Nondestructive estimation of standing crop and fuel moisture content in tallgrass prairie. *Rangel. Ecol. Manag.* **2018**, *71*, 356–362. [[CrossRef](#)]
54. Zhang, X.; Chen, B.; Zhao, H.; Fan, H.; Zhu, D. Soil moisture retrieval over a semiarid area by means of PCA dimensionality reduction. *Can. J. Remote Sens.* **2016**, *42*, 136–144. [[CrossRef](#)]
55. Pasolli, L.; Notarnicola, C.; Bruzzone, L.; Bertoldi, G.; Della Chiesa, S.; Niedrist, G.; Tappeiner, U.; Zebisch, M. Polarimetric RADARSAT-2 imagery for soil moisture retrieval in alpine areas. *Can. J. Remote Sens.* **2012**, *37*, 535–547. [[CrossRef](#)]
56. Liu, D.; Mishra, A.K.; Yu, Z. Evaluating uncertainties in multi-layer soil moisture estimation with support vector machines and ensemble Kalman filtering. *J. Hydrol.* **2016**, *538*, 243–255. [[CrossRef](#)]
57. Lee, Y.; Jung, C.; Kim, S. Spatial distribution of soil moisture estimates using a multiple linear regression model and Korean geostationary satellite (COMS) data. *Agric. Water Manag.* **2019**, *213*, 580–593. [[CrossRef](#)]
58. Bortolini, D.; Albuquerque, J.A. Estimation of the retention and availability of water in soils of the State of Santa Catarina. *Rev. Bras. Ciência Do Solo* **2018**, *42*, 1–13. [[CrossRef](#)]
59. Carranza, C.; Nolet, C.; Pezij, M.; Ploeg, M. Van Der Root zone soil moisture estimation with Random Forest. *J. Hydrol.* **2021**, *593*, 125840. [[CrossRef](#)]
60. Gupta, D.K.; Prasad, R.; Kumar, P.; Vishwakarma, A.K. Soil moisture retrieval using ground based bistatic scatterometer data at X-band. *Adv. Space Res.* **2017**, *59*, 996–1007. [[CrossRef](#)]
61. Chakravorty, A.; Chahar, B.R.; Sharma, O.P.; Dhanya, C.T. A regional scale performance evaluation of SMOS and ESA-CCI soil moisture products over India with simulated soil moisture from MERRA-Land. *Remote Sens. Environ.* **2016**, *186*, 514–527. [[CrossRef](#)]
62. Leng, P.; Song, X.; Li, Z.L.; Ma, J.; Zhou, F.; Li, S. Bare surface soil moisture retrieval from the synergistic use of optical and thermal infrared data. *Int. J. Remote Sens.* **2014**, *35*, 988–1003. [[CrossRef](#)]
63. Liu, M.; Huang, C.; Wang, L.; Zhang, Y.; Luo, X. Short-term soil moisture forecasting via Gaussian process regression with sample selection. *Water* **2020**, *12*, 3085. [[CrossRef](#)]
64. Wang, Y.; Shi, L.; Xu, T.; Zhang, Q.; Ye, M.; Zha, Y. A nonparametric sequential data assimilation scheme for soil moisture flow. *J. Hydrol.* **2021**, *593*, 125865. [[CrossRef](#)]
65. Xu, L.; Wang, Q. Retrieval of soil water content in saline soils from emitted thermal infrared spectra using partial linear squares regression. *Remote Sens.* **2015**, *7*, 14646–14662. [[CrossRef](#)]
66. Nakamura, K.; Yasutaka, T.; Kuwatani, T.; Komai, T. Development of a predictive model for lead, cadmium and fluorine soil-water partition coefficients using sparse multiple linear regression analysis. *Chemosphere* **2017**, *186*, 501–509. [[CrossRef](#)] [[PubMed](#)]
67. Qiu, Y.; Fu, B.; Wang, J.; Chen, L. Spatiotemporal prediction of soil moisture content using multiple-linear regression in a small catchment of the Loess Plateau, China. *Catena* **2003**, *54*, 173–195. [[CrossRef](#)]
68. Chen, Z.; Wang, X. Model for estimation of total nitrogen content in sandalwood leaves based on nonlinear mixed effects and dummy variables using multispectral images. *Chemom. Intell. Lab. Syst.* **2019**, *195*, 103874. [[CrossRef](#)]
69. Qiu, Y.; Fu, B.; Wang, J.; Chen, L.; Meng, Q.; Zhang, Y. Spatial prediction of soil moisture content using multiple-linear regressions in a gully catchment of the Loess Plateau, China. *J. Arid Environ.* **2010**, *74*, 208–220. [[CrossRef](#)]
70. Soleimani, R.; Chavoshi, E.; Shirani, H.; Pour, I.E. Comparison of stepwise multilinear regressions, artificial neural network, and genetic algorithm-based neural network for prediction the plant available water of unsaturated soils in a semi-arid region of Iran (case study: Chaharmahal Bakhtiari province). *Commun. Soil Sci. Plant. Anal.* **2020**, *51*, 2297–2309. [[CrossRef](#)]
71. Pahlavan-rad, M.R.; Dahmardeh, K.; Hadizadeh, M. Prediction of soil water infiltration using multiple linear regression and random forest in a dry flood plain, eastern Iran. *Catena* **2020**, *194*, 104715. [[CrossRef](#)]
72. Mahmoud, M.A.; El, A.; Saad, N.; Shaer, R. El Phase II multiple linear regression profile with small sample size. *Qual. Reliab. Eng. Int.* **2015**, *31*, 851–861. [[CrossRef](#)]
73. Jenkins, D.G.; Quintana-Ascencio, P.F. A solution to minimum sample size for regressions. *PLoS ONE* **2020**, *15*, e0229345. [[CrossRef](#)] [[PubMed](#)]

74. Zhao, C.; Jia, X.; Zhu, Y.; Shao, M. Long-term temporal variations of soil water content under different vegetation types in the Loess Plateau, China. *Catena* **2017**, *158*, 55–62. [CrossRef]
75. Su, C.; Fu, B.-J. Evolution of ecosystem services in the Chinese Loess Plateau under climatic and land use changes. *Glob. Planet. Chang.* **2013**, *101*, 119–128. [CrossRef]
76. Zhao, J.; van Oost, K.; Chen, L.; Govers, G. Moderate topsoil erosion rates constrain the magnitude of the erosion-induced carbon sink and agricultural productivity losses on the Chinese Loess Plateau. *Biogeosciences* **2016**, *13*, 4735–4750. [CrossRef]
77. Xin, Z.; Yu, X.; Li, Q.; Lu, X.X. Spatiotemporal variation in rainfall erosivity on the Chinese Loess Plateau during the period 1956–2008. *Reg. Environ. Chang.* **2011**, *11*, 149–159. [CrossRef]
78. Tasumi, M.; Kimura, R. Estimation of volumetric soil water content over the Liudaogou river basin of the Loess Plateau using the SWEST method with spatial and temporal variability. *Agric. Water Manag.* **2013**, *118*, 1–7. [CrossRef]
79. Hu, W.; Shao, M.; Han, F.; Reichardt, K. Spatio-temporal variability behavior of land surface soil water content in shrub- and grass-land. *Geoderma* **2011**, *162*, 260–272. [CrossRef]
80. Chen, J.; Wang, C.; Jiang, H.; Mao, L.; Yu, Z. Estimating soil moisture using temperature-vegetation dryness index (TVDI) in the Huang-huai-hai (HHH) plain. *Int. J. Remote Sens.* **2011**, *32*, 1165–1177. [CrossRef]
81. He, J.; Yang, X.H.; Huang, S.F.; Di, C.L.; Mei, Y. Study on soil moisture by thermal infrared data. *Therm. Sci.* **2013**, *17*, 1375–1381. [CrossRef]
82. Yang, R.W.; Wang, H.; Hu, J.M.; Cao, J.; Yang, Y. An improved temperature vegetation dryness index (iTVDI) and its applicability to drought monitoring. *J. Mt. Sci.* **2017**, *14*, 2284–2294. [CrossRef]
83. Claps, P.; Laguardia, G. Assessing spatial variability of soil water content through thermal inertia and NDVI. *Remote Sens. Agric. Ecosyst. Hydrol. V* **2004**, *5232*, 378. [CrossRef]
84. Price, J.C. On the analysis of thermal infrared imagery: The limited utility of apparent thermal inertia. *Remote Sens. Environ.* **1985**, *18*, 59–73. [CrossRef]
85. Capodici, F.; Cammalleri, C.; Francipane, A.; Ciralo, G.; la Loggia, G.; Maltese, A. Soil water content diachronic mapping: An FFT frequency analysis of a temperature–vegetation index. *Geoscience* **2020**, *10*, 23. [CrossRef]
86. Dong, J.; Steele-Dunne, S.C.; Judge, J.; van de Giesen, N. A particle batch smoother for soil moisture estimation using soil temperature observations. *Adv. Water Resour.* **2015**, *83*, 111–122. [CrossRef]
87. Cohen, A. Dummy variables in stepwise regression. *Am. Stat.* **1991**, *45*, 226–228. [CrossRef]
88. Cox, N.J.; Schechter, C.B. Speaking stata: How best to generate indicator or dummy variables. *Stata J.* **2019**, *19*, 246–259. [CrossRef]
89. Li, L.; Lien, B.; Carmen, S.; Frank, C.; Kervyn, M. Dating lava flows of tropical volcanoes by means of spatial modeling of vegetation recovery. *Earth Surf. Process. Landf.* **2018**, *43*, 840–856. [CrossRef]
90. Chen, M.; Zhang, Y.; Yao, Y.; Lu, J.; Pu, X.; Hu, T.; Wang, P. Evaluation of the OPTRAM model to retrieve soil moisture in the Sanjiang Plain of northeast China. *Earth Space Sci.* **2020**, *7*. [CrossRef]
91. Level-1 and Atmosphere Archive and Distribution System (LAADS) Distributed Archive Center (DAAC). Available online: <https://ladsweb.modaps.eosdis.nasa.gov/> (accessed on 21 February 2019).
92. Chen, T.; de Jeu, R.A.M.; Liu, Y.Y.; van der Werf, G.R.; Dolman, A.J. Using satellite based soil moisture to quantify the water driven variability in NDVI: A case study over mainland Australia. *Remote Sens. Environ.* **2014**, *140*, 330–338. [CrossRef]
93. Wagle, P.; Xiao, X.; Torn, M.S.; Cook, D.R.; Matamala, R.; Fischer, M.L.; Jin, C.; Dong, J.; Biradar, C. Sensitivity of vegetation indices and gross primary production of tallgrass prairie to severe drought. *Remote Sens. Environ.* **2014**, *152*, 1–14. [CrossRef]
94. Sharma, S.; Carlson, J.D.; Krueger, E.S.; Engle, D.M.; Twidwell, D.; Fuhlendorf, S.D.; Patrignani, A.; Feng, L.; Ochsner, T.E. Soil moisture as an indicator of growing-season herbaceous fuel moisture and curing rate in grasslands. *Int. J. Wildland Fire* **2020**, *30*, 57–69. [CrossRef]
95. Wang, S.; Mo, X.; Liu, S.; Lin, Z.; Hu, S. Validation and trend analysis of ECV soil moisture data on cropland in North China Plain during 1981–2010. *Int. J. Appl. Earth Obs. Geoinf.* **2016**, *48*, 110–121. [CrossRef]
96. McNally, A.; Shukla, S.; Arsenault, K.R.; Wang, S.; Peters-Lidard, C.D.; Verdin, J.P. Evaluating ESA CCI soil moisture in East Africa. *Int. J. Appl. Earth Obs. Geoinf.* **2016**, *48*, 96–109. [CrossRef] [PubMed]
97. Xin, Q.; Li, J.; Li, Z.; Li, Y.; Zhou, X. Evaluations and comparisons of rule-based and machine-learning-based methods to retrieve satellite-based vegetation phenology using MODIS and USA National Phenology Network data. *Int. J. Appl. Earth Obs. Geoinf.* **2020**, *93*, 102189. [CrossRef]
98. Li, L.; Zhou, X.; Chen, L.; Chen, L.; Zhang, Y.; Liu, Y. Estimating urban vegetation biomass from Sentinel-2A image data. *Forests* **2020**, *11*, 125. [CrossRef]
99. Yang, X.; Li, L.; Chen, L.; Chen, L.; Shen, Z. Improving ASTER GDEM accuracy using land use-based linear regression methods: A case study of Lianyungang, East China. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 145. [CrossRef]
100. Awange, J.L.; Gebremichael, M.; Forootan, E.; Wakbulcho, G.; Anyah, R.; Ferreira, V.G.; Alemayehu, T. Characterization of Ethiopian mega hydrogeological regimes using GRACE, TRMM and GLDAS datasets. *Adv. Water Resour.* **2014**, *74*, 64–78. [CrossRef]
101. Yu, B.; Liu, G.; Liu, Q.; Wang, X.; Feng, J.; Huang, C. Soil moisture variations at different topographic domains and land use types in the semi-arid Loess Plateau, China. *Catena* **2018**, *165*, 125–132. [CrossRef]
102. Geng, R.; Zhang, G.H.; Ma, Q.H.; Wang, H. Effects of landscape positions on soil resistance to rill erosion in a small catchment on the Loess Plateau. *Biosyst. Eng.* **2017**, *160*, 95–108. [CrossRef]

103. Panciera, R.; Walker, J.P.; Kalma, J.; Kim, E. A proposed extension to the soil moisture and ocean salinity level 2 algorithm for mixed forest and moderate vegetation pixels. *Remote Sens. Environ.* **2011**, *115*, 3343–3354. [[CrossRef](#)]
104. Raoult, N.; Delorme, B.; Ottlé, C.; Peylin, P.; Bastrikov, V.; Maugis, P.; Polcher, J. Confronting soil moisture dynamics from the ORCHIDEE land surface model with the ESA-CCI product: Perspectives for data assimilation. *Remote Sens.* **2018**, *10*, 1786. [[CrossRef](#)]
105. Sun, A.Y.; Xia, Y.; Caldwell, T.G.; Hao, Z. Patterns of precipitation and soil moisture extremes in Texas, US: A complex network analysis. *Adv. Water Resour.* **2018**, *112*, 203–213. [[CrossRef](#)]
106. Huza, J.; Teuling, A.J.; Braud, I.; Grazioli, J.; Melsen, L.A.; Nord, G.; Raupach, T.H.; Uijlenhoet, R. Precipitation, soil moisture and runoff variability in a small river catchment (Ardeche, France) during HyMeX Special Observation Period 1. *J. Hydrol.* **2014**, *516*, 330–342. [[CrossRef](#)]
107. China Meteorological Data Service Center. Available online: <http://data.cma.cn/en> (accessed on 11 January 2019).
108. Cenci, L.; Pulvirenti, L.; Boni, G.; Pierdicca, N. Defining a trade-off between spatial and temporal resolution of a geosynchronous SAR mission for soil moisture monitoring. *Remote Sens.* **2018**, *10*, 1950. [[CrossRef](#)]
109. Cheng, L.; Li, L.; Chen, L.; Hu, S.; Yuan, L.; Liu, Y.; Cui, Y.; Zhang, T. Spatiotemporal variability and influencing factors of Aerosol Optical Depth over the Pan Yangtze River Delta during the 2014–2017 period. *Int. J. Environ. Res. Public Health* **2019**, *16*, 3522. [[CrossRef](#)]
110. Wang, S.; Fu, B.; Gao, G.; Liu, Y.; Zhou, J. Responses of soil moisture in different land cover types to rainfall events in a re-vegetation catchment area of the Loess Plateau, China. *Catena* **2013**, *101*, 122–128. [[CrossRef](#)]
111. Wang, X.; Wang, B.; Xu, X.; Liu, T.; Duan, Y.; Zhao, Y. Spatial and temporal variations in surface soil moisture and vegetation cover in the Loess Plateau from 2000 to 2015. *Ecol. Indic.* **2018**, *95*, 320–330. [[CrossRef](#)]
112. Sharma, M.J.; Yu, S.J. Stepwise regression data envelopment analysis for variable reduction. *Appl. Math. Comput.* **2015**, *253*, 126–134. [[CrossRef](#)]
113. Eeftens, M.; Beelen, R.; de Hoogh, K.; Bellander, T.; Cesaroni, G.; Cirach, M.; Declercq, C.; Dedele, A.; Dons, E.; de Nazelle, A.; et al. Development of land use regression models for PM<sub>2.5</sub>, PM<sub>2.5</sub> absorbance, PM<sub>10</sub> and PM<sub>coarse</sub> in 20 European study areas; Results of the ESCAPE project. *Environ. Sci. Technol.* **2012**, *46*, 11195–11205. [[CrossRef](#)]
114. Hirsch-Eshkol, T.; Baharad, A.; Alpert, P. Investigation of the dominant factors influencing the ERA15 temperature increments at the subtropical and temperate belts with a focus over the Eastern Mediterranean Region. *Land* **2014**, *3*, 1015–1036. [[CrossRef](#)]
115. Lewis-Beck, M.; Bryman, A.; Futing Liao, T. Stepwise Regression. In *SAGE Encyclopedia of Social Science Research Methods*; SAGE: London, UK, 2012; pp. 1–9.
116. Ebrahimi-Khusfi, M.; Alavipanah, S.K.; Hamzeh, S.; Amiraslani, F.; Neysani Samany, N.; Wigneron, J.P. Comparison of soil moisture retrieval algorithms based on the synergy between SMAP and SMOS-IC. *Int. J. Appl. Earth Obs. Geoinf.* **2018**, *67*, 148–160. [[CrossRef](#)]
117. Wang, Z.X.; Zhang, H.L.; Zheng, H.H. Estimation of Lorenz curves based on dummy variable regression. *Econ. Lett.* **2019**, *177*, 69–75. [[CrossRef](#)]
118. Holgersson, T.; Nordström, L.; Öner, Ö. On regression modelling with dummy variables versus separate regressions per group: Comment on Holgersson et al. *J. Appl. Stat.* **2016**, *43*, 1564–1565. [[CrossRef](#)]
119. Chen, D.; Huang, X.; Zhang, S.; Sun, X. Biomass modeling of larch (*Larix* spp.) plantations in China based on the mixed model, dummy variable model, and Bayesian hierarchical model. *Forests* **2017**, *8*, 268. [[CrossRef](#)]
120. Jiao, Q.; Li, R.; Wang, F.; Mu, X.; Li, P.; An, C. Impacts of re-vegetation on surface soil moisture over the Chinese Loess Plateau based on remote sensing datasets. *Remote Sens.* **2016**, *8*, 156. [[CrossRef](#)]
121. Colliander, A.; Fisher, J.B.; Halverson, G.; Merlin, O.; Misra, S.; Bindlish, R.; Jackson, T.J.; Yueh, S. Spatial downscaling of SMAP soil moisture using MODIS land surface temperature and NDVI during SMAPVEX15. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 2107–2111. [[CrossRef](#)]
122. Brust, C.; Kimball, J.S.; Maneta, M.P.; Jencso, K.; He, M.; Reichle, R.H. Using SMAP Level-4 soil moisture to constrain MOD16 evapotranspiration over the contiguous USA. *Remote Sens. Environ.* **2021**, *255*, 112277. [[CrossRef](#)]
123. Pan, J.; Bai, Z.; Cao, Y.; Zhou, W.; Wang, J. Influence of soil physical properties and vegetation coverage at different slope aspects in a reclaimed dump. *Environ. Sci. Pollut. Res.* **2017**, *24*, 23953–23965. [[CrossRef](#)]
124. Xu, C.; Qu, J.J.; Hao, X.; Zhu, Z.; Gutenberg, L. Surface soil temperature seasonal variation estimation in a forested area using combined satellite observations and in-situ measurements. *Int. J. Appl. Earth Obs. Geoinf.* **2020**, *91*, 102156. [[CrossRef](#)]
125. Wu, Z.; Lei, S.; Bian, Z.; Huang, J.; Zhang, Y. Study of the desertification index based on the albedo-MSAVI feature space for semi-arid steppe region. *Environ. Earth Sci.* **2019**, *78*, 1–13. [[CrossRef](#)]