https://doi.org/10.5194/essd-2024-200
Preprint. Discussion started: 15 July 2024

Open Access Earth System Science Data Discussions

GSSM: A global seamless soil moisture dataset from 1981 to 2022 matching CCI to SMAP with a novel bias correction method

Yunjia Wang[1], Hao Sun[1], Zhenheng Xu[1], Jinhua Gao[1], Huanyu Xu[1], Tian Zhang[1], Dan Wu[1,2]

[1]College of Geoscience and Surveying Engineering, China University of Mining and Technology-Beijing, Beijing, 100083, China

[2]Remote Sensing Research Institute of Ningxia, Yinchuan 750021, China

Correspondence to: Hao Sun (sunhao@cumtb.edu.cn)

Abstract. Surface soil moisture is vital for Earth's environmental and energy cycles. However, it is still rare to have remote sensing soil moisture data with a long-term temporal extent, a global seamless spatial coverage, and a near-real-time update frequency. Here, we provided a global seamless soil moisture dataset from July 1981 to December 2022, matching CCI with SMAP through a novel soil moisture data bias correction method (fitting beta CDF matching, BCDF), and filling the gaps of corrected soil moisture through XGBoost Algorithms along with various soil moisture covariates. The new soil moisture dataset was abbreviated as GSSM and it has been validated with in situ observations, original CCI and SMAP data, and simulated gap areas. Results demonstrated that 1) the GSSM has similar accuracy with the SMAP and they are both more accurate than the original CCI data as compared with in situ observations at 399 global sites (averaged R=0.72, averaged ubRMSE<0.05); 2) the GSSM has the global spatial coverage, while filling the gaps of original CCI data through various soil moisture covariates (in artificial gaps verification, averaged R>0.86, averaged ubRMSE<0.04); 3) the GSSM has the same temporal variation characteristics with the original CCI dataset, while it can be combined with SMAP to obtain a long-term and near-real-time soil moisture dataset. Thus, GSSM provides long-term and seamless soil moisture data, paving the way for environmental disaster and water cycle process research.

1 Introduction

Surface soil moisture, also known as surface soil water content, plays a vital role in environmental water cycle processes and energy transfer processes in Earth's surface systems (Green et al., 2019; Gianotti et al., 2019; Vereecken et al., 2008; Babaeian et al., 2019). It is also regarded as an important climate indicator by the Global Climate Observing System (GCOS) (Al-Yaari et al., 2017). Beyond this, soil moisture data are needed in monitoring agricultural droughts(Pan et al., 2023), floods worldwide(Chen et al., 2023), water resourece management (Robinson et al., 2008), and climate change (Anderson et al., 2007).

With the deepening of global climate change research, a global seamless, long-term, and near-real-time soil moisture data has become more and more important. From a temporal perspective, long-term soil moisture data are needed to analyze seasonal and long-term changes in soil moisture accurately. This kind of data can be used not only to analyze the impact of climate

https://doi.org/10.5194/essd-2024-200
Preprint. Discussion started: 15 July 2024

change on soil moisture (Shellito et al., 2016), but also to evaluate the frequency and duration of drought and wet cycles (Sheffield and Wood, 2008), and to study the relationship between soil moisture and vegetation growth, the relationship between agricultural production and ecosystem health (Bertoldi et al., 2016). From a spatial perspective, seamless soil moisture data with global coverage are needed to compare and monitor soil moisture conditions in different regions, such as studying soil moisture climate changes in tropical rainforest regions (Ma et al., 2023). In terms of accuracy, high-quality soil moisture data is needed to ensure accuracy and reliability, thereby supporting various soil moisture applications in agricultural management, water resources management, and climate research. For example, SMAP has limited product error to less than 0.04 m^3/m^3 in many validation and evaluation studies conducted at global and regional scales (Chan et al., 2016; Colliander et al., 2017; Yao et al., 2021), which can better understand processes that link the terrestrial water, energy, and carbon cycles (Bai et al., 2019; Entekhabi et al., 2010). Therefore, taking into account the requirements of time, space, and accuracy, higher requirements are put forward for the acquisition and processing of soil moisture data. So, how to obtain soil moisture data that integrates wide spatial coverage, long time range, and high accuracy?

Currently, there are three methods to obtain high-accuracy soil moisture data with global seamless spatial characteristics and long-term, near-real-time time characteristics: traditional ground-based measurements at monitoring stations, reanalysis products, and remote sensing techniques. The method of obtaining soil moisture through ground stations has the characteristics of high precision, temporal continuity, and excellent data quality. However, it is limited to point-scale measurements, which is affected by site density distribution and makes real-time monitoring expensive (Rahimzadeh-Bajgiran et al., 2013). The second is reanalyzing soil moisture products simulated through a meteorological model. Soil moisture reanalysis data can break through the limitations of satellite-borne signal-derived data, achieve full coverage of soil moisture, and have clear physical meanings. (Liu et al., 2023). It possesses characteristics of broad spatiotemporal coverage and relatively high precision. Reanalysis products have become essential for providing continuous soil moisture data over large areas. The quality of these products varies despite their comprehensive consideration of factors and coverage of various meteorological data. These products predict temporal changes well, but the bias and root mean square error (RMSE) can be significant (Bi et al., 2016). Since the 1980s, microwave remote sensing data for spatially and temporally continuous operations over large areas has become an attractive option for drought monitoring, especially when ground measurements are impossible (Sadri et al., 2020). Nowadays, microwave remote sensing has become the leading method for soil moisture estimation due to its ability to penetrate clouds and vegetation while obtaining data in near-real-time (Karthikeyan et al., 2017). Compared with the first two methods, remote sensing technology has become the most promising way to obtain soil moisture data in long-term series, near-real-time, and high spatial coverage.

https://doi.org/10.5194/essd-2024-200
Preprint. Discussion started: 15 July 2024

Table 1: Basic information on currently available global remotely sensed soil moisture datasets.

Product	Spatial resolution	Temporal resolution	Temporal extent
ESA CCI	0.25°	One day	1978-2022
AMSR-E	25 km	Two-three day	2002-2011
AMSR-2	25 km	Two-three day	2012- Now
SMOS	50 km	Two-three day	2010- Now
SMAP	36 km/ 9km	Two-three day	2015-Now
FY-3C	25 km	One day	2014-2020

Nowadays, there are many global soil moisture data sets based on remote sensing (for example, shown in Table 1). Different soil moisture datasets have different characteristics and applicable scopes. The update frequency of most remote sensing soil moisture data can be updated in near-real-time. Nevertheless, the temporal extent of most remote sensing soil moisture datasets is limited, influencing their applicability for long-term soil moisture time series analysis (Escorihuela and Quintana-Seguí, 2016; Ford and Quiring, 2019). For example, the soil moisture products in Table 1 can all achieve global coverage, but the time series of SMAP, SMOS, and FY-3C are relatively short and only available after 2010. Although some data are very long in time series, their accuracy performance is not ideal. For example, the accuracy of AMSR-E/AMSR-2 is more prone to errors and biases compared with SMAP SSM products in the interaction of atmosphere, vegetation, and soil (Yao et al., 2021). From the perspective of climate change research, CCI data makes up for the above shortcomings. It has the longest time series, global coverage, and daily temporal resolution. Despite the extensive temporal span of the CCI soil moisture dataset, limitations remain. First, the update frequency is irregular, which affects the near-real-time availability of data. Secondly, its large amount of missing data limits comprehensive coverage and affects the effectiveness of soil moisture monitoring. CCI datasets are severely missing globally, especially in mainland China. The average ratio of missing data to the total data volume is around 40%, and in winter and spring, its proportion can reach up to 80% (Sun and Xu, 2021). Furthermore, the lack of data makes it challenging to maintain spatial continuity of CCI soil moisture data (Llamas et al., 2020). At the same time, compared with SMAP, after comparing various remote sensing soil moisture data with ground measured data, it was found that the accuracy of SMAP soil moisture products is better than that of CCI and is closest to the measured data, and SMAP data have the potential to be integrated into existing long-term ESA CCI products to form a more reliable and useful product (Ma et al., 2019; Kim et al., 2018; Kumar et al., 2018; Cui et al., 2018). To sum up, the shortcomings of CCI data are reflected in data update frequency, data spatial coverage, and data accuracy. Nowadays, there are currently few soil moisture remote sensing products that can simultaneously provide span long-time series, higher spatial coverage, and high data accuracy.

Fortunately, the above characteristics can be achieved through the fusion of multiple datasets and gap filling (González-Zamora et al., 2019). SM products with higher spatial coverage can be obtained through filling methods, and long-term, near-real-time, high-accuracy products can be obtained through data fusion methods. The previous research has solved the problem of low spatial coverage. The current mainstream method is to use machine learning or deep learning methods to fill in soil moisture

https://doi.org/10.5194/essd-2024-200
Preprint. Discussion started: 15 July 2024

data. Zhang et al. (2022) integrated data from three sensors, namely AMSR-E, AMSR2, and WindSat, and employed a long short-term memory convolutional neural network (LSTM-CNN) to interpolate soil moisture data, achieving favorable outcomes. Sun et al. (2023) used geographical information and meteorological or climate factors as filled SM covariates, selected the XGBoost model to fill in the CCI products of the Chinese region from 1982 to 2020, and obtained seamless long-time series CCI products of the Chinese region. However, it is limited to filling in the mainland China area and does not achieve global coverage. At the same time, data fusion is used to solve the near-real-time and long-term problems of CCI data. Since there are systematic errors in different soil moisture products (such as errors caused by different sensors and different inversion algorithms), the two products cannot be directly fused, but an appropriate assimilation method needs to be used (Su et al., 2013; Lee et al., 2017; Konings et al., 2011). Using appropriate fusion methods can not only expand the time series of soil moisture products, but also improve product accuracy. At present, data fusion methods can be divided into linear methods and non-linear methods. Nonlinear methods are commonly used for data fusion, among which Cumulative Distribution Function Mapping (CDFM) and machine learning methods are the most widely used (Kornelsen and Coulibaly, 2015; Afshar and Yilmaz, 2017). For example, Sadri et al. (2020) used CDFM and Bayesian conditional process methods, combining SMAP with SMOS to obtain near-real-time global soil moisture with an accuracy similar to CCI products. Yao et al. (2023) used artificial neural networks to fuse the SMAP dataset and the long-term brightness temperature data of the FY-3B satellite to develop an SM dataset from 2010 to 2019, whose accuracy is close to that of SMAP. Yang et al. (2024) extended the SMAP dataset with the corresponding CCI SM time series by using a random forest model with an accuracy close to that of the SMAP product. However, in predicting long-term trends in geoscience variables, machine learning methods are severely challenged by factors such as limited historical data, the non-stationary nature of geoscience processes (cyclones and floods) (Karpatne et al., 2019). The CDF method can avoid the above problems well, so the CDF matching method still has research potential (Ji et al., 2020). However, the CDF matching method also has the problem of how to determine the boundary value.

In order to solve the above problems, we use a novel matching method (BCDF) to determine boundary values, apply gap filling methods (XGBoost) using various geoscientific covariates to the global scale, and propose a long-term, seamless, high-accuracy soil moisture dataset called GSSM. It has high accuracy, long time series, high spatial coverage, and near-real-time capabilities that can be combined with SMAP. The dataset follows a unified latitude and longitude grid, with a spatial resolution of 0.25° × 0.25° and a monthly temporal resolution. Detailed matching and filling methods, as well as dataset verification methods, will be systematically elaborated in Section 2. Section 3 will focus on the verification results of the GSSM dataset. In Section 4, we will discuss matching algorithms, strategies for determining boundary values, and the application details of freeze-thaw masks.

https://doi.org/10.5194/essd-2024-200
Preprint. Discussion started: 15 July 2024

2 Methods and materials

2.1 Datasets

2.1.1 ESA CCI

The Soil Moisture CCI Combined dataset is one of three datasets formulated within the framework of the European Space Agency's (ESA) Soil Moisture Essential Climate Variable (ECV) Climate Change Initiative (CCI) project. Its products are created by directly merging scatterometers (active remote sensing) and radiometers (passive remote sensing) derived from multiple satellites (Dorigo et al., 2017; Preimesberger et al., 2021; Gruber et al., 2019). The CCI V08.1 data was updated on October 11, 2023, with a temporal resolution of one day and a spatial resolution of 0.25°. The time span is from November 1, 1978, to December 31, 2022, with a total of 16,132 images. In Wang et al. (2023) research, compared with the soil moisture of a single satellite data, the combined CCI data has higher precision, so the combined CCI product was selected for our research.

2.1.2 SMAP

Since January 31, 2015, the Soil Moisture Active Passive (SMAP) satellite equipped with an L-band radiometer and an L-band radar has been observing the Earth in a sun-synchronous orbit. The satellite passes over the Equator at 06:00 (descending) and 18:00 (ascending) during its orbit (Entekhabi et al., 2010). In our study, we selected the SPL3SMP_E v005 ascending orbit data. This product has a spatial resolution of 9km and a temporal resolution of two-three day. It is an enhanced level 3 soil moisture product that provides soil moisture active and passive radiometers retrieved. A synthesis of daily estimates of global surface conditions.

2.1.3 Other data

According to the filling algorithm of Sun et al. (2023), several geographical information and meteorological or climate factors as filled SM covariates have been applied to fill the product, including ERA5-Land, GIMMS/MOD13C2, HWSD v2.0, GTOPO30 DEM. The reason why ERA5-Land was chosen to fill in the soil moisture data is that, among the products evaluated, ERA5-Land always performed better, showed a preferable ability to capture spatial and temporal changes in SM, and had a higher correlation with ISMN (Zhang et al., 2023). All data sources are shown in Table 1. We pass each data through Monthly Fusion and then resample each data to 0.25°.

Table 2 Geospatial and meteorological information for gap filling.

Variables	Data	Time Range	Temporal resolution	Spatial resolution	Data availability (URL)
NDVI	GIMMS	1981.07-2015.12	15day	0.083°	https://ecocast.arc.nasa.gov/data/pub/gimms/

https://doi.org/10.5194/essd-2024-200
Preprint. Discussion started: 15 July 2024

	MOD13C2	2000.02-Now	Monthly	0.05°	https://ladsweb.modaps.eosdis.nasa.gov/missions-and-measurements/products/MOD13C2
Background Soil Moisture	Volumetric soil water layer 1	1950.01-Now	Monthly	0.1°	https://www.ecmwf.int/en/era5-land
Albedo	Forecast albedo	1950.01 -now	Monthly	0.1°	
Surface Temperature	Soil temperature level 1	1950.01-Now	Monthly	0.1°	
Air Temperature	2m temperature	1950.01-Now	Monthly	0.1°	
Precipitation	Total precipitation	1950.01-Now	Monthly	0.1°	
Potential Evapotranspiration	Total evaporation	1950.01-Now	Monthly	0.1°	
Soil Texture	HWSD v2.0	-	-	0.083°	https://www.fao.org/soils-portal/data-hub/soil-maps-and-databases/harmonized-world-soil-database-v20/en/
DEM	GTOPO30	-	-	0.083°	https://www.usgs.gov/centers/eros/science/usgs-eros-archive-digital-elevation-global-30-arc-second-elevation-gtopo30

2.1.4 Validation data

In situ measurements from the International Soil Moisture Network (ISMN) (Dorigo et al., 2011; Dorigo et al., 2013). We choose ISMN field measurement data as the actual measured value to verify the product accuracy after matching and filling. We selected a total of 24 detection networks on ISMN for accuracy verification. Since GSSM represents the surface soil moisture in the range of 0-5cm, the ISMN site data of 0-5cm was selected for verification. Due to the variability in quality among ISMN in situ, we have formulated selection rules for ISMN site data: (1) The length of soil moisture in situ data recorded exceeds one year, and the length of the time series of the soil moisture product at the pixel position is not less than one year; (2) Since there are missing pixels in the dataset, only the data where the site position time series and the dataset

https://doi.org/10.5194/essd-2024-200
Preprint. Discussion started: 15 July 2024

corresponding pixel position time series exist at the same time are selected; (3) Only the quality flag is selected as G (GOOD) the in situ is verified; (4) Only validation data with dated site data is selected for comparison. According to the above rules, a total of 24 site networks and 399 site data meets the requirements. Fig. 1 illustrates the spatial distribution of the selected station data in our study.

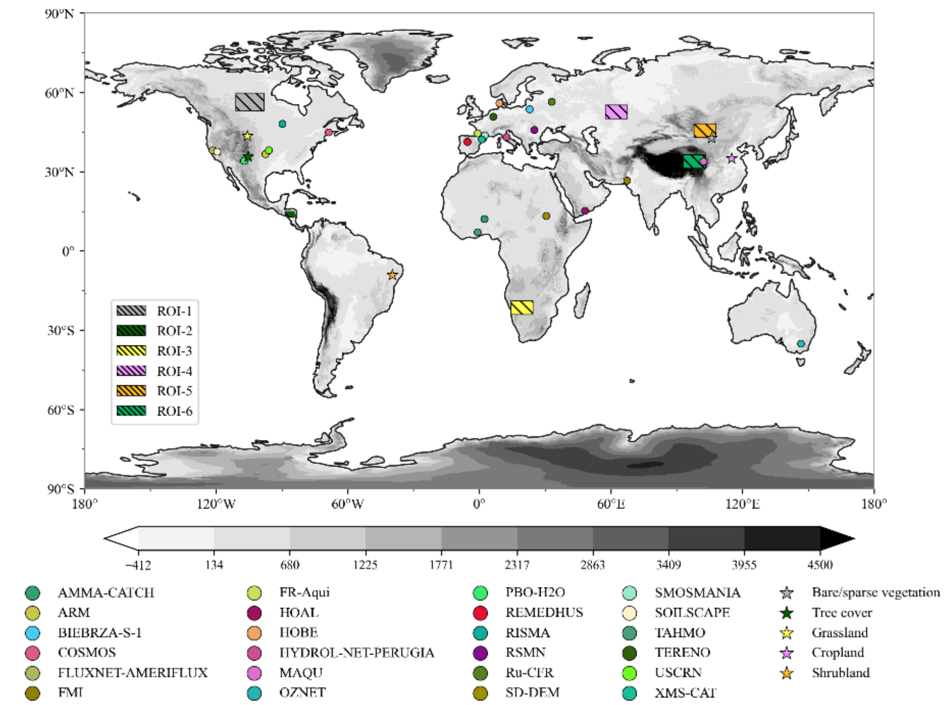

Figure 1: Global distribution of networks and sites in the ISMN dataset used in our study, along with a schematic representation of the location of typical land cover types and areas with significant dry and wet changes (Digital elevation model (DEM) represented by base map).

2.2 Methodology

The production of a global long-term series, and gap filling surface soil moisture dataset consists of four basic parts. (1) pre-processing, including data selecting, resampling, and monthly average synthesis of various data; (2) deviation correction to match the ESA CCI data to the SMAP dataset through the fitted beta distribution CDF method, and generate the corresponding relationship through the overlap time of SMAP and CCI (2015.03.31-2022.12.31), and apply this correspondence to obtain the SMAP-corrected daily and monthly CCI/SMAP datasets from 1978.11.01-2022.12.31; (3) gap filling, use the XGBoost method to fill in some areas that can be filled in the CCI/SMAP dataset, and obtain GSSM dataset; (4) post-processing, freeze-thaw masking is performed on the filled soil moisture data set to mask out areas with freeze water and null values. The overall methodological framework for producing a long-time, and gap-filling 0.25° GSSM product is shown in Fig. 2, with details described in the following context of this section.

https://doi.org/10.5194/essd-2024-200
Preprint. Discussion started: 15 July 2024

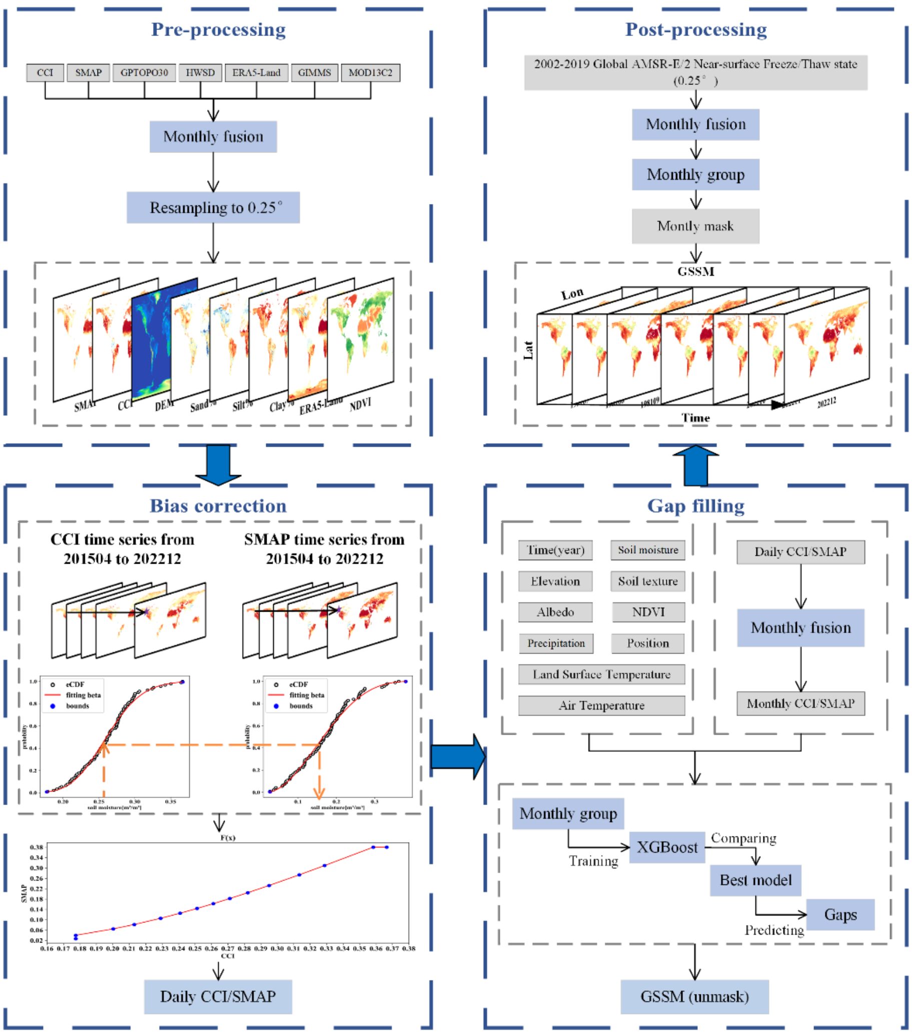

Figure 2: The overall methodological framework of our study.

2.2.1 Bias correction method for the production of global long-term surface soil moisture data

Cumulative distribution function matching (CDFM) can be considered a way to reduce the systematic differences between source and reference datasets (Reichle and Koster, 2004; Crow and Van Den Berg, 2010; Draper et al., 2011). CDF matching was applied for each grid point individually. The CDF is a specific way to give the probability that X will take a value less than or equal to a certain threshold (Madelon et al., 2022).

$$CDF_{SM}(\mathrm{X}) = \mathrm{P}(SM \le X)\ , \tag{1}$$

There are multiple methods to match the CDFs of two datasets, such as linear piecewise interpolation and polynomial fitting. Besides, there is a linear method that directly corrects for bias between mean and variance. We assume that the CDF of soil moisture for each grid point matches a beta distribution for modelling soil moisture time series (Reichle and Koster, 2004). In

https://doi.org/10.5194/essd-2024-200
Preprint. Discussion started: 15 July 2024

the study by Sadri et al. (2018), several parameter distributions (including normal distribution and Gumbel distribution) were used to fit the soil moisture time series, and it was found that the beta distribution showed the best goodness of fit. The general formula for the beta probability density function (pdf) is:

$$f(x) = \frac{(x-a)^{(p-1)}(b-x)^{(q-1)}}{B(p,q)(b-a)^{p+q-1}}, a \le x \le b, p, q > 0, \quad B(p,q) = \int_0^1 t^{p-1}(1-t)^{q-1}\,\mathrm{d}t. \tag{2}$$

185 Where p, q is the shape parameter of beta distribution; a, b is the upper and lower bounds, which we will call the boundary later. When $a = 0$, $b = 1$, it is called the standard beta distribution. Where $B(p, q)$ is the beta constant calculated from the above formula. Therefore, we performed beta distribution fitting on the time series of soil moisture at each pixel position of CCI and SMAP, and used the moment of moments to fit the beta distribution (Reichle and Koster, 2004). In our research, the difference is that we adopt a novel method suitable for our study in selecting the boundary value: for each pixel's time series, after fitting it to a beta distribution, the minimum and maximum observations in the data set are compared to the minimum and maximum values of the percentile point function (ppf), respectively, and the data are sorted in ascending order, to achieve the purpose of determining boundaries. The actual algorithm is shown below Eq. (3).

$$TS_{SM}^{a} = [Min\left(ppf_{SM}^{a}(0), Min(SM^{a})\right), ppf_{SM}^{a}(0), \dots, ppf_{SM}^{a}(1), Max(ppf_{SM}^{a}(1), Max(SM^{a}))], \tag{3}$$

Where SM stands for the soil moisture data from both the CCI and SMAP; TS represents the time series of SM at pixel position a; $ppf(0)$denotes the SM corresponding to the minimum quantile of the dataset after fitting it to a beta distribution; $ppf(1)$ denotes the soil moisture value corresponding to the maximum quantile of the dataset after fitting it to a beta distribution; $Min(,)$ means taking the minimum value of the two; $Max(,)$ means taking the maximum value of the two. Obtain the corresponding CDF distribution after fitting, and perform CDF matching on CCI and SMAP.

After testing, the overall correction accuracy of the fitting beta CDF matching (BCDF) method is slightly higher than that of LR, the direct CDF segment matching method, and the CDF fitting method (Discussion 4.1).

Due to the existence of standard deviation in the calculation formula, the matching method is not available when there is only one value in the time series. Therefore, for these "special" pixels, we adopt the nearest neighbour interpolation method for correction, that is, select the nearest neighbour correspondence to correct the pixel. A correspondence was established based on the overlapping period of SMAP and CCI data (April 2015 to December 2022), which was then used to extrapolate the SMAP-corrected CCI/SMAP dataset for the period spanning November 1, 1978, to December 31, 2022.

2.2.2 Gap Filling method for the production of global gap-filling surface soil moisture data

As only the values of CCI are subjected to bias correction, the corrected CCI/SMAP product still exhibits gaps, thereby posing limitations in long-term soil drought research. Consequently, it is necessary to fill gaps in SSM product. We referred to Sun et

https://doi.org/10.5194/essd-2024-200
Preprint. Discussion started: 15 July 2024

al. (2023) gap filling method based on machine learning (ML) and used the XGBoost model to fill gaps. The principal formula is such as Eq. (4) and Eq. (5).

$$SM^{a}_{CCI/SMAP} = f^{a}(Time^{a}, Position^{a}, Elevation^{a}, Soil\ texture^{a}, Meteorological\ factors^{a})\ , \tag{4}$$

$$SM^{g}_{Predict} = f^{a}(Time^{g}, Position^{g}, Elevation^{g}, Soil\ texture^{g}, Meteorological\ factors^{g})\ , \tag{5}$$

a represents the available SM pixel position; g represents the gap SM pixel position; $SM^{a}_{CCI/SMAP}$ refers to the CCI/SMAP soil moisture value at the "a" pixel position; f^{a} means a filling model obtained through machine learning training; $Time^{a}$, etc. represent the filling features at the "a" pixel position; $Time^{g}$, etc. represent the filling features at the "g" pixel position; $SM^{g}_{Predict}$ refers to the soil moisture value predicted by the model at the "g" pixel position. The principle of machine learning is to build a model through machine learning methods based on the available SM and various SM covariates, and then use the specified model for the available SM covariates to estimate the SM of the gap, so as to achieve the purpose of filling (Sun et al., 2023).

Regarding the covariates for filling SSM, in previous studies(Sun and Cui, 2021; Sun and Xu, 2021; Sun et al., 2023), geographical information and climate factors were used. Hence, we chose to include Normalized Differential Vegetation Index (NDVI), Albedo (A), Land Surface Temperature (LST), Air Temperature (AT), Precipitation (P), Potential Evapotranspiration (PET), Soil Texture (ST), Elevation (DEM), background SM from ERA5-Land, and time information (year). The reason for choosing ERA5-Land to fill in the soil moisture data is that among the evaluated products, ERA5-Land consistently exhibits superior performance, demonstrating a strong capability to capture spatial and temporal variations in soil moisture. It also shows a higher correlation with ISMN (Zhang et al., 2023).

2.2.3 Methods for the validation of surface soil moisture products

In order to comprehensively evaluate the matching and filling effects, we choose four indicators to evaluate product quality, including correlation coefficient (R) as Eq. 6, average bias(Bias) as Eq. 7, root mean square error (RMSE) as Eq. 8 and ubRMSE as Eq. 8 (Sun and Cui, 2021; Kornelsen and Coulibaly, 2015).

$$R = \frac{\sum(\theta_o - E[\theta_o])(\theta_r - E[\theta_r])}{\sqrt{\sum(\theta_o - E[\theta_o])^2 \sum(\theta_r - E[\theta_r])^2}}\ , \tag{6}$$

$$Bias = \mathrm{E}[\sum(\theta_o - \theta_r)]\ , \tag{7}$$

$$RMSE = \sqrt{E[(\theta_o - \theta_r)^2]}\ , \tag{8}$$

$$ubRMSE = \sqrt{E[\big((\theta_o - E[\theta_o]) - (\theta_r - E[\theta_r])\big)^2]}\ , \tag{9}$$

Where $E(,)$ refers to take the mean of the data in brackets; θ_o, θ_r represent the corrected or predicted soil moisture value and the reference soil moisture value.

The following three verification methods are used for the bias correction results: (1) verification by comparison with SMAP time series; (2) verification by comparison with SMAP data in space and time; (3) in situ verification. Use SMAP data in the time range of 2015.03.31-2022.12.31 to verify CCI/SMAP products. We roughly selected six areas with obvious dry and wet changes according to Fig. 1 (Liu et al., 2023). The purpose is to test the accuracy of the product in terms of time and space. At the same time, ISMN is used to verify the dataset to see whether the dataset meets the SSM accuracy requirements.

The following two verification methods are used for the filling results: (1) simulated missing area verification; (2) simulated in situ verification. Six areas with obvious dry and wet changes were excavated, and the filling model was used to predict them. The purpose was to test the prediction accuracy of the prediction model in time and space. At the same time, when evaluating the filling precision, we compared and verified the SM obtained by ISMN in situ observation with the filled dataset to verify the overall accuracy of the filling product.

3 Validation

3.1 The spatiotemporal distribution of the GSSM dataset

By employing the BCDF correction method, we brought the CCI data closer to SMAP in terms of numerical values and obtained corresponding monthly GSSM products. Subsequently, leveraging various auxiliary datasets and employing the XGBoost machine learning method, we filled the gaps in the GSSM monthly products. The filling process spanned from July 1981 to December 2022, resulting in a nearly 42-year seamless soil moisture dataset. Numerical restrictions are applied to the filling to prevent soil moisture values that exceed the actual physical meaning. The restricted moisture value is between 0.02 and 1. Fig. 3 illustrates the comparison of BCDF-corrected GSSM soil moisture before and after filling over several months spanning 40 years (1981.12, 1990.11, 1998.10, 2006.9, 2014.8, 2022.7). Comparing the images before and after filling in Fig. 3, we can see that the soil moisture product before filling has spatial discontinuities in the CCI, so the corrected data still has such characteristics. In spring and winter, there is a serious lack of data in high-latitude areas, such as Russia and some European countries. After filling in the BCDF-corrected GSSM soil moisture data from July 1981 to December 2022, the integrity of the spatial data has been greatly improved. Compared with the soil moisture data before filling, The filled spatial data distribution is more continuous and almost complete in space.

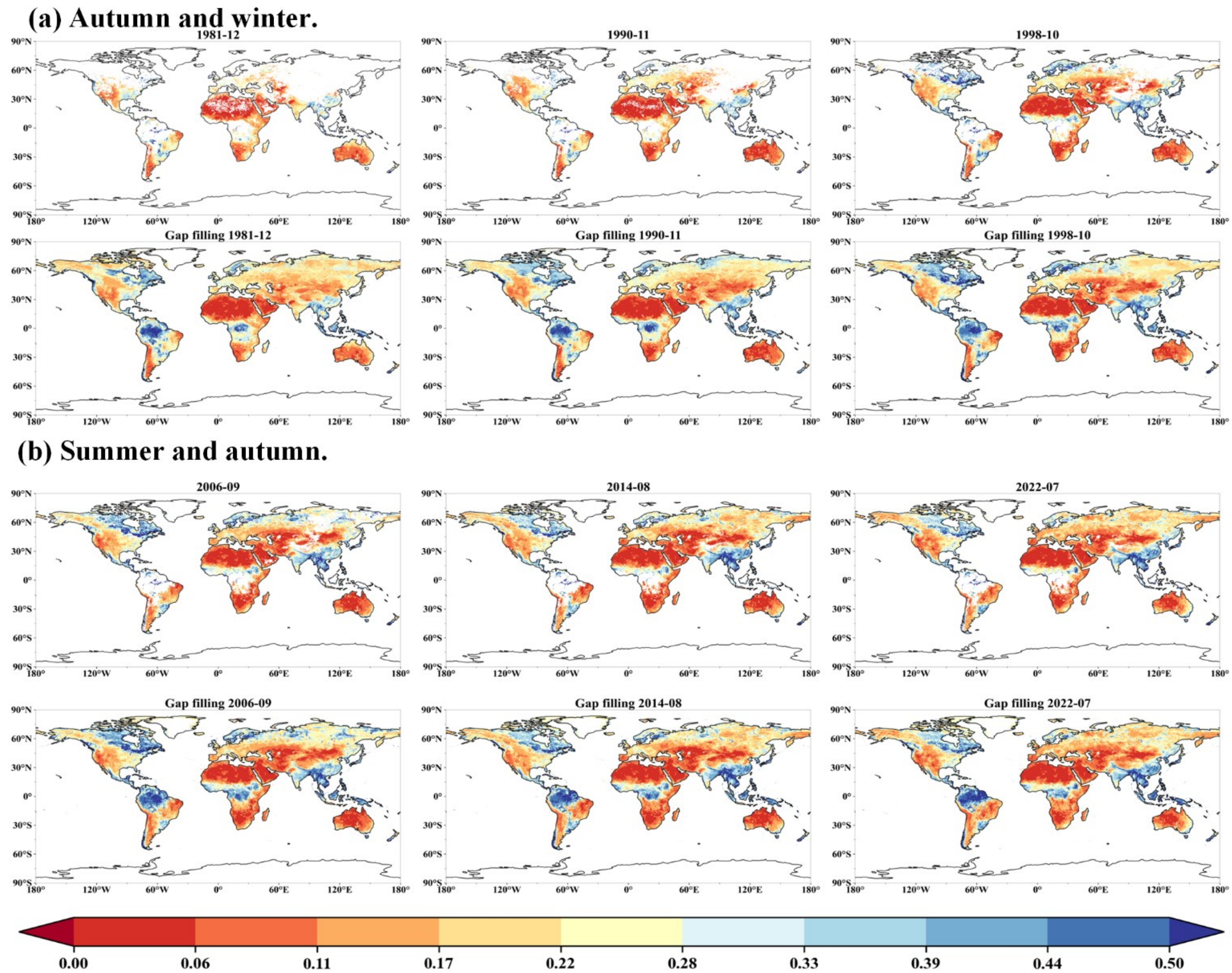

Figure 3: Spatial comparison of soil moisture dataset before and after gap filling. (a)The first line is the soil moisture image before partial date filling in autumn and winter, and the second line is the image after filling. (b)The first line is the soil moisture image before partial date filling in summer and autumn, and the second line is the image after filling.

We selected five land cover categories based on the land cover product ESAWorldCover10m v200, and extracted the soil moisture time series at the pixel locations of the five land cover types (Bare/spare vegetation, Tree cover, Grassland, Cropland, Shrubland). The specific location information is shown in Table 3 and Fig. 1. Comparative analysis was conducted on the original CCI, BCDF-corrected and filled CCI, and SMAP soil moisture data time series, and the results in Fig. 4 were obtained. Overall, the CCI soil moisture sequence after BCDF gap filling is closer to the SMAP soil moisture time series, with great performance in terms of precision. The original CCI and SMAP soil moisture time series in Fig. 4(a) are discontinuous in time, which is not conducive to long-term series analysis of soil moisture data. Meanwhile, the CCI soil moisture series after BCDF

https://doi.org/10.5194/essd-2024-200
Preprint. Discussion started: 15 July 2024

corrected and gap filling is not only numerically closer to SMAP but also has increased the time continuity to increase the length of time that soil moisture data can be used.

Table 3 Basic information about typical features.

Index	Lon	Lat	Main land use
01	105.83	42.74	Bare/spare vegetation
02	-105.52	35.59	Tree cover
03	-105.74	43.46	Grassland
04	115.07	35.06	Cropland
05	-39.69	-9.06	Shrubland

275

Figure 4: Correction effect on typical land cover type time series. (a)-(e) shows the soil moisture time series corresponding to five land cover categories.

https://doi.org/10.5194/essd-2024-200
Preprint. Discussion started: 15 July 2024

3.2 Evaluation of GSSM with SMAP products

When evaluating the matching precision, we selected SMAP data and ISMN datasets to perform an accuracy analysis of the matched daily GSSM product to test the accuracy of the corrected CCI product. SMAP data is used to verify whether the corrected product has reduced the gap with SMAP data, thereby verifying whether it can be combined with SMAP data to achieve the purpose of near-real-time. We resampled the SMAP product to 0.25° resolution for comparison with the CCI and daily GSSM datasets. At the same time, all ascending orbit data of the SPL3SMP_E v005 product on the NASA website were selected, with a total of 2748 images, covering the period from March 31, 2015 to December 31, 2022, which is consistent with the GSSM dataset.

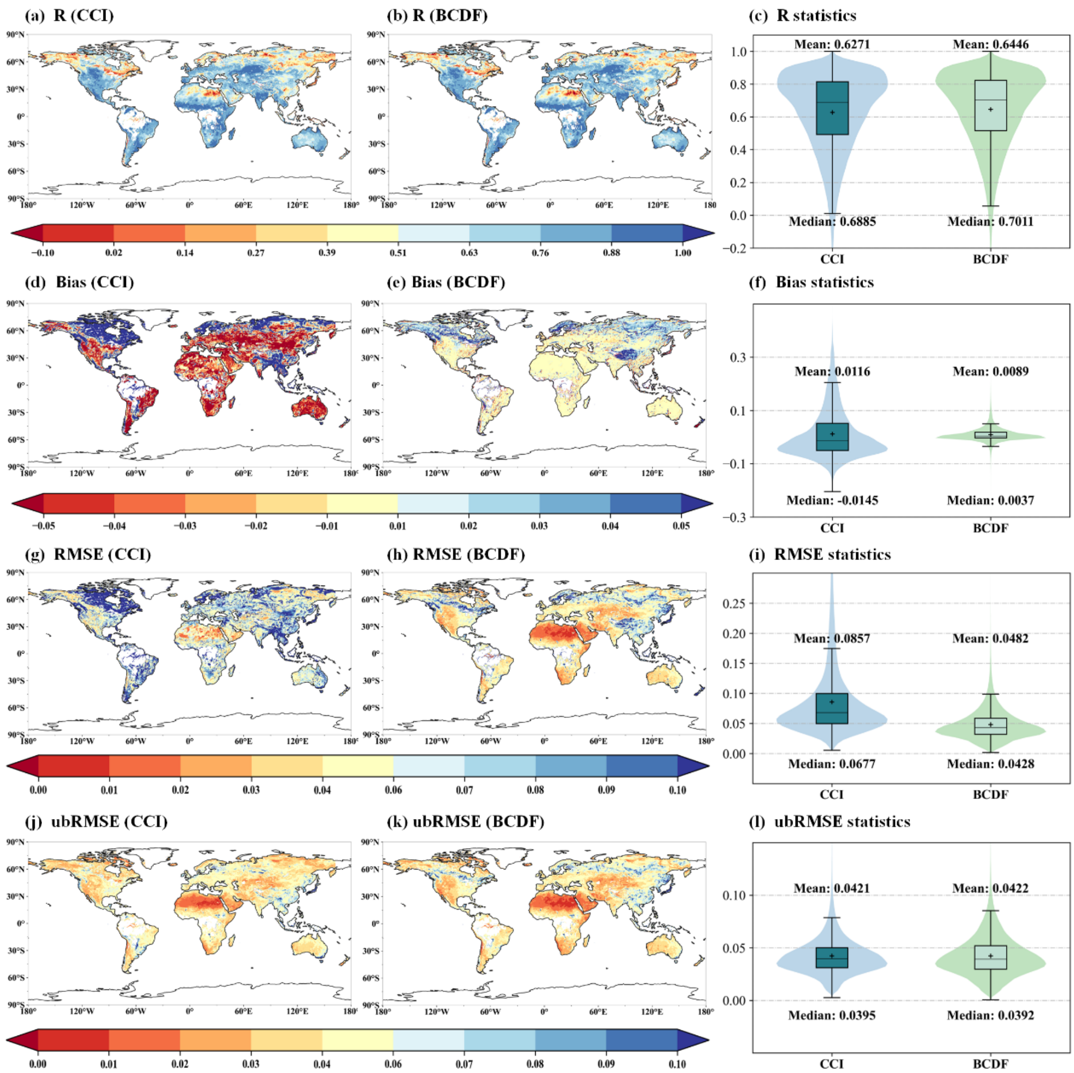

Figure 5: Comparing the correlation (a) and (b), bias (d) and (e), RMSE (g) and (h), as well as ubRMSE (g) and (h) between the original CCI product and the CCI product corrected using CDFM, and SMAP. The period of comparison is from 2015/03/31 to 2022/12/31.

https://doi.org/10.5194/essd-2024-200
Preprint. Discussion started: 15 July 2024

In Fig. 5, a comparison of the R (correlation), bias, RMSE (Root Mean Square Error), and ubRMSE (unbiased Root Mean Square Error) are presented among the ESA CCI product, the CCI product corrected using BCDF, and the SMAP product. The GSSM dataset obtained from BCDF method in column 2 (Fig. 5b, Fig. 5c, Fig. 5h, Fig. 5k) reveals lower RMSE and Bias with SMAP globally, compared to the dataset obtained from the origin dataset in column 1 (Fig. 5a, Fig. 5d, Fig. 5g, Fig. 5j). From Fig. 5c, the overall correlation changes before and after correction is not obvious. This may be because CDF has the advantage of maintaining the variation characteristics of the original time series (Cui et al., 2018). Therefore, the original temporal variation characteristics of CCI are retained. Before the correction, the overall accuracy of CCI data was lower than that of SMAP, and the deviation was larger in high latitudes. After correction, the average bias from SMAP was significantly reduced, especially the bias in high latitudes was also corrected to a relatively small range. RMSE is significantly lower than before correction, with significant improvements in northern Africa, southern North America and the Middle East. The ubRMSE exhibits consistent performance before and after correction. Although the correlation coefficient does not change significantly in the overall image, compared with before correction, R has improved numerically, and Bias and RMSE have significantly decreased. A comprehensive evaluation of the matching effect was carried out based on R, RMSE, Bias, and ubRMSE indicators. The most obvious performance was in correlation, bias, and RMSE. The correlation increased by 0.0175, an increase of about 3%; the average Bias and RMSE decreased by 0.0027cm³/cm³ and 0.0375cm³/cm³, which are reduced by about 23% and 44%, indicating satisfactory matching performance. The above results show that the BCDF matching method we proposed can effectively reduce the systematic error between CCI and SMAP products while retaining the temporal variation characteristics of the original data.

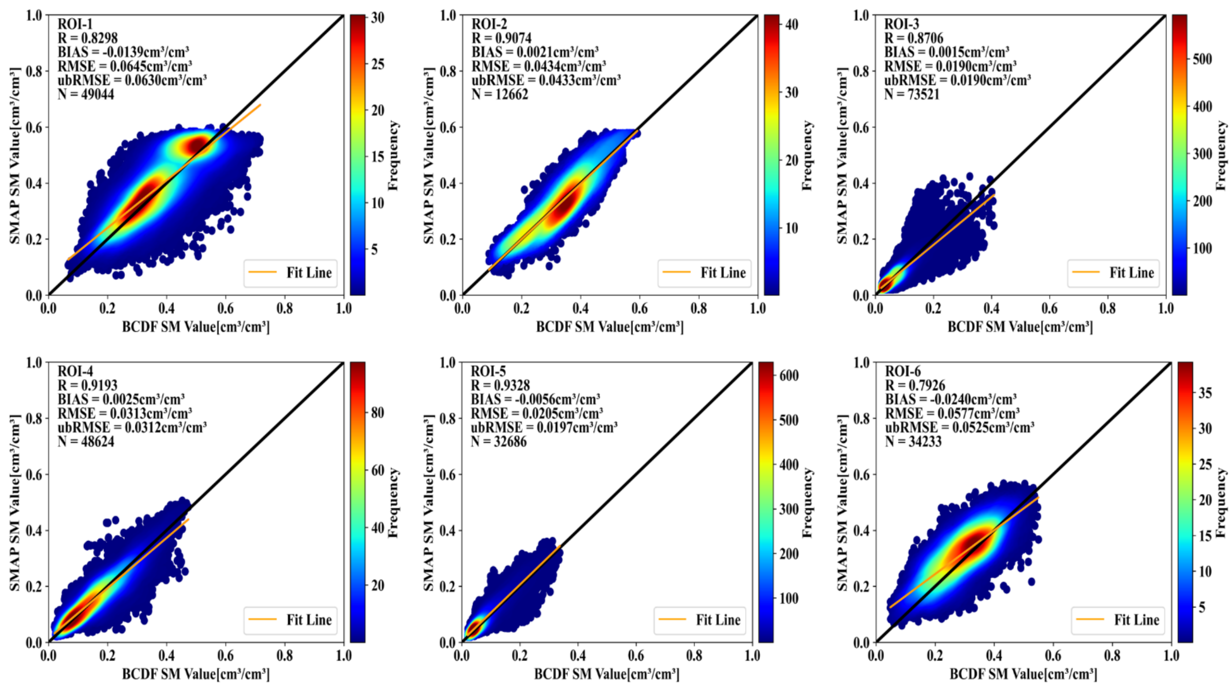

Figure 6: Spatiotemporal analysis results before and after matching in six selected areas.

https://doi.org/10.5194/essd-2024-200
Preprint. Discussion started: 15 July 2024

We conducted a spatiotemporal analysis (from April 2015 to December 2022) on the six selected areas to verify the accuracy effect between the corrected CCI and SMAP. The results are shown in Fig. 6. It can be seen from the six verification results that the CCI data and SMAP data after BCDF matching have achieved relatively close performance. Across the six regions, the average correlation coefficient exceeds 0.88, with average Bias, RMSE, and ubRMSE of -0.0062 cm^3/cm^3, 0.0394 cm^3/cm^3, and 0.0381 cm^3/cm^3, respectively. The high correlation and low Bias, RMSE, and ubRMSE demonstrate the strong consistency, both numerically and spatially, between the BCDF-corrected CCI data and SMAP data.

3.3 Evaluation of GSSM with in situ observations

We selected the data with data quality "G" on the ISMN website, a total of 24 site networks (a total of 399 sites), and used the 24 site network data as verification data. We compared the original CCI data and the BCDF-corrected CC with SMAP data, and the overall accuracy verification results are shown in Fig. 7. We noticed that there are some negative values in the correlation. This may be because due to different scales, SMAP and CCI reflect macro-scale soil moisture conditions compared with the point-scale, there are differences in soil moisture values, resulting in a negative correlation. However, since there is temporal stability, that is, local-scale ground soil moisture can still reflect the temporal dynamics of large-area soil moisture, we chose this method to verify the accuracy(Brocca et al., 2009). In general, SMAP data shows a closer alignment with ground station measurements than the original CCI data. Compared with the measured in situ, the SMAP data has a lower Bias, RMSE, and ubRMSE, which further proves the rationality of CCI matching to SMAP. On the whole, the CCI data after BCDF matching has improved in all four inspection indicators. Compared to the original CCI, the BCDF-corrected exhibits an increase of 0.0007 in the correlation with the mean of the station data, and decreases in Bias, RMSE, and ubRMSE by 0.0307, 0.0107, and 0.0021, respectively. Based on the four evaluation indicators, compared with the in situ, the accuracy of the corrected GSSM dataset is close to that of the SMAP product and better than that of the CCI product that has not been corrected by BCDF.

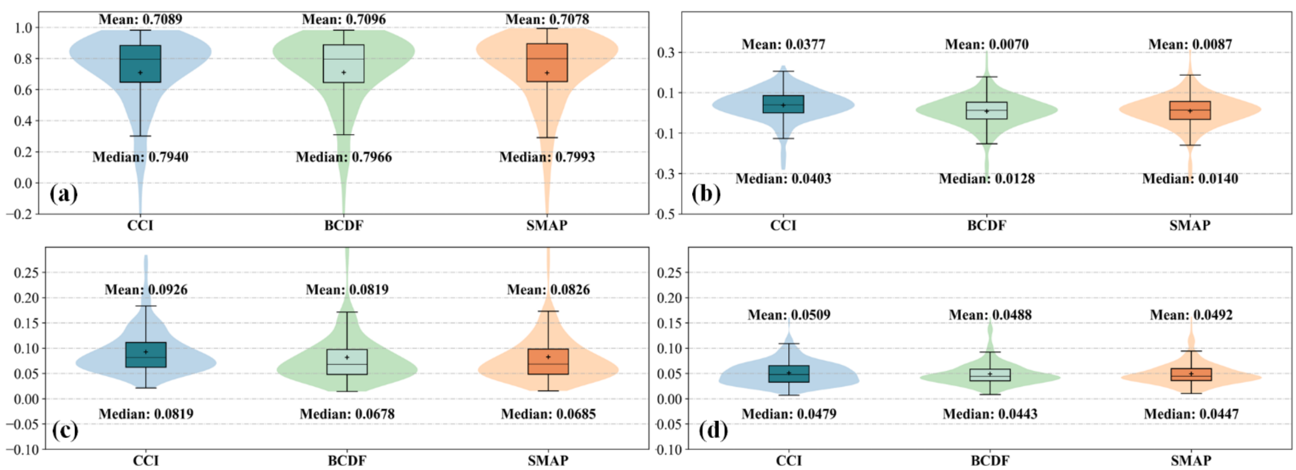

Figure 7: Metrics of R, Bias, RMSE, ubRMSE. Displayed from left to right are the correlation comparisons between CCI products, BCDF-corrected CCI products, and SMAP products with measured soil moisture in situ.

https://doi.org/10.5194/essd-2024-200
Preprint. Discussion started: 15 July 2024

3.4 Evaluation of GSSM with simulated SM gaps

We used a training set and test set to verify the fitting effect of XGBoost. The soil moisture data of six regions of interest (Fig. 1) distributed within the study area were removed from the training set, resulting in six artificial gaps. Apply the gap-filling method to these gaps, that is, use XGBoost to predict the values in these areas and then compare the predicted value (Predicted Value) with the data in these areas in the GSSM data (Original Value) as Fig. 8. Across the six regions, XGBoost has excellent accuracy in filling the area, which is better than the accuracy requirement of SMAP (average ubRMSE<0.04). The average correlation exceeds 0.86, with mean biases, RMSE, and ubRMSE of -0.0005 cm³/cm³, 0.0394 cm³/cm³, and 0.0380 cm³/cm³, respectively. It can be seen from the data that the predicted results have lower Bias, RMSE and ubRMSE, and higher R, which shows that XGBoost performs well in predicting soil moisture data.

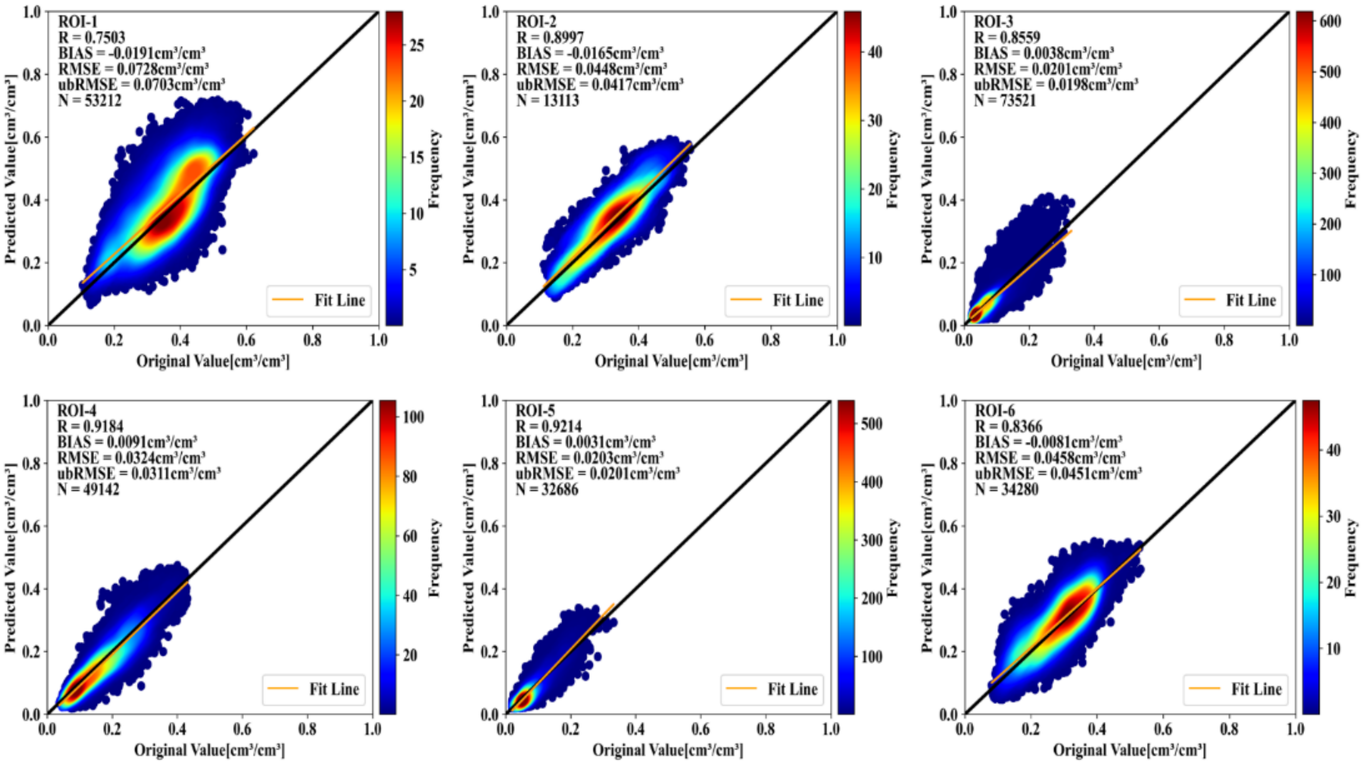

Figure 8: XGBoost prediction effect at six artificial gaps.

We filled the data corrected by the BCDF method, compared the filled results with the measured site data, and obtained the results in Fig.9. Judging from the verification results, the correlation has improved after filling, increasing by 0.0065. And the bias, RMSE, and ubRMSE have also improved. Statistics show that XGBoost padded data has superior precision. The overall accuracy is close to the SMAP data and better than the original CCI data.

https://doi.org/10.5194/essd-2024-200
Preprint. Discussion started: 15 July 2024

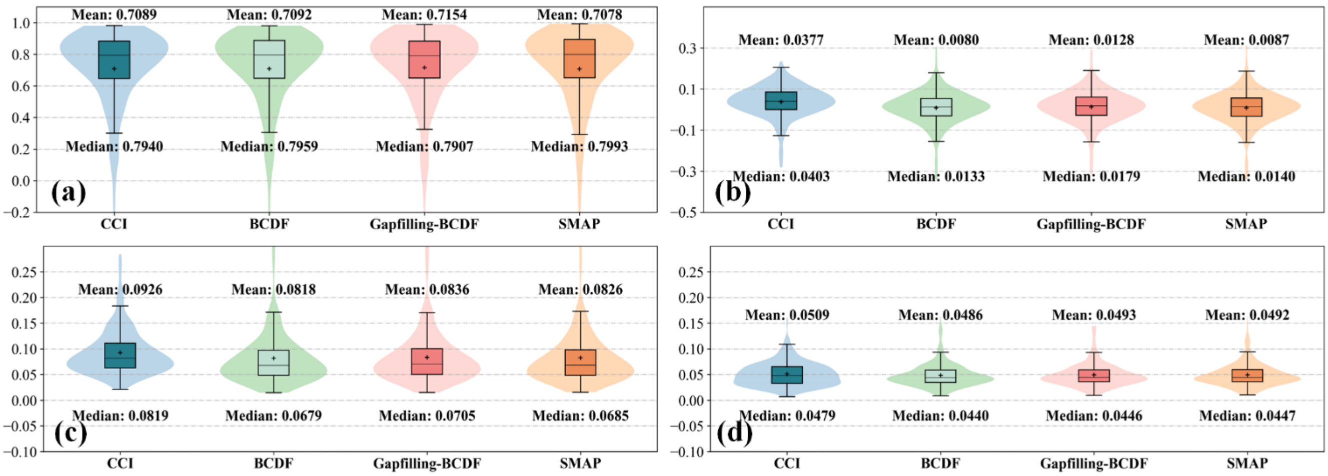

Figure 9: Accuracy verification results obtained by comparing the filled results with the measured in situ.

4 Discussion

4.1 Comparison and validation of bias matching methods

Before bias correction, we selected the more mainstream linear method linear rescaling (LR) (Draper et al., 2009) and the nonlinear method piecewise linear CDF (LCDF) (Liu et al., 2011; Reichle and Koster, 2004; Drusch et al., 2005), fitting polynomial CDF (MCDF) (Aires et al., 2021; Brocca et al., 2011; Madelon et al., 2022) method and our own proposed fitting beta distribution CDF (BCDF) method. Among them, the piecewise linear CDF matching method is the method currently used by the ESA CCI project (Moesinger et al., 2020). Accuracy verification of all data corrections from April 2015 to December 2022.

Compared to the randomly selected time series of five land types (Fig. 1), Table 4 shows the performance of various methods at each location. Based on the results in Table 4, there will be two situations. On the one hand, the nonlinear method CDF accuracy is better than the linear LR method. On the other hand, the linear method LR is better than most nonlinear CDF methods. The BCDF bias correction method demonstrated excellent accuracy performance in both cases, with multiple accuracy indicators outperforming other methods in each case. In the statistical results of correction indicators (Table 4), BCDF performs relatively well in each statistical indicator. We found that LCDF and MCDF methods sometimes reduce correlation, while LR and BCDF methods both improve correlation. It is worth noting that in terms of comprehensive accuracy evaluation, BCDF performs better in accuracy indicators in most cases, with higher R and lower Bias, RMSE, and ubRMSE.

https://doi.org/10.5194/essd-2024-200
Preprint. Discussion started: 15 July 2024

Table 4 Various bias correction methods match the statistical results of simulations (Bold font indicates the best performing indicator among each matching method.).

Land covers	Method	R	Bias	RMSE	ubRMSE
Bare/sparse vegetation	CCI	0.7989	0.0796	0.0809	0.0147
	LR	0.7989	-0.0051	0.0143	0.0133
	LCDF	0.7790	-0.0132	0.0192	0.0140
	MCDF	0.8253	**0.0000**	**0.0129**	0.0129
	BCDF	**0.8287**	-0.0046	0.0134	**0.0126**
Grassland	CCI	0.8716	0.0425	0.0480	0.0222
	LR	0.8716	-0.0007	0.0218	0.0218
	LCDF	0.8066	-0.0099	0.0278	0.0260
	MCDF	0.8662	**0.0000**	0.0225	0.0225
	BCDF	**0.8738**	-0.0007	**0.0216**	**0.0216**
Shrubland	CCI	0.8842	0.0487	0.0573	0.0302
	LR	0.8842	0.0018	0.0232	0.0232
	LCDF	0.7611	0.0386	0.0696	0.0580
	MCDF	0.8865	**0.0000**	0.0233	0.0233
	BCDF	**0.8875**	0.0018	**0.0229**	**0.0229**
Tree cover	CCI	0.7104	0.0867	0.1016	**0.0530**
	LR	0.7104	-0.0041	0.0566	0.0565
	LCDF	0.7128	-0.0113	0.0553	0.0541
	MCDF	0.7157	**0.0000**	0.0550	0.0550
	BCDF	**0.7252**	-0.0044	**0.0549**	0.0547
Cropland	CCI	0.8866	0.0721	0.0761	0.0243
	LR	0.8866	0.0000	0.0143	0.0143
	LCDF	0.8777	0.0000	0.0150	0.0150
	MCDF	0.8734	0.0000	0.0151	0.0151
	BCDF	**0.8871**	0.0000	**0.0143**	**0.0143**

4.2 Bias correction method boundary determination

The time series of soil moisture is consistent with the beta distribution, and the beta distribution is determined by the shape parameters and location parameters. When we fit the beta distribution, its data range can be extended. Due to the flexibility of the beta distribution, it enables the establishment of more extensive data relationships in bias correction, thereby achieving

https://doi.org/10.5194/essd-2024-200
Preprint. Discussion started: 15 July 2024

superior performance in the correction process. Therefore, when performing BCDF matching, it is necessary to determine the boundaries in the fitting parameters. How should we determine the boundaries? The method adopted by Sheffield et al. (2004) is to sort the data, take the sum of the top 10% and the bottom 10% for linear fitting, and extrapolate to estimate the lower limit and upper limit. Abourizk et al. (1994) suggestion is to choose the maximum value of the data as the boundary value. In our experiments, however, these two methods did not yield satisfactory results. Hence, we propose a novel method for boundary determination. The approach we adopted involves fitting the time series of each pixel to a beta distribution, comparing the minimum and maximum observed values in the dataset with the extremes of ppf data, reordering the data in ascending order, and thereby determining the boundary values. The actual algorithm is shown in Eq. 3.

After literature research and combined with the actual practice of this experiment, we tested three methods, namely linear regression interpolation of data as boundary values, direct selection of boundary values, and our newly proposed method to determine the boundary. The corresponding distributions obtained by the three methods are shown in Fig. 10. Method-3 represents the method we proposed, which can effectively expand the boundary values of soil moisture on both sides. Based on the time series analysis of the pixel positions, the boundary values obtained by the first boundary determination method do not fully cover the entire time series, so there will be frequent outliers during the correction process, especially in areas with low soil moisture values. However, it extends the distribution on the right side of the soil moisture data to a certain extent, but there are problems with the extension effect. In the third pixel's position, we can see that the boundary determined by this method is too extended, so that it appears on the image showing a nearly parallel trend. The second and third methods can all extend the boundaries on both sides, but the third method can extend it more effectively than the second method, which is reflected in a larger numerical range. To sum up, the third method we proposed achieves a more effective effect of extending the boundary value.

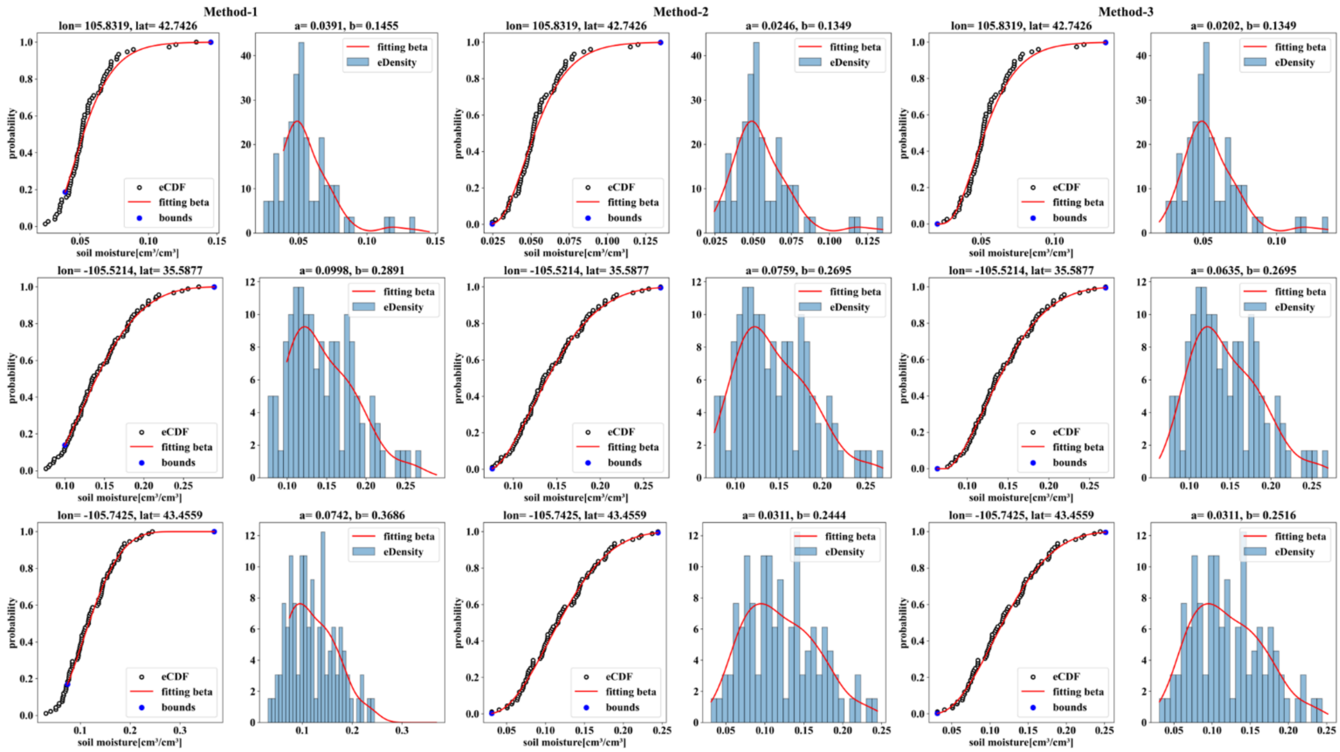

Figure 10: Data distribution CDF and PDF results obtained by three methods.

4.3 Determination of active water pixel position

Due to limitations in data acquisition, the means used to verify the accuracy of soil moisture in our study were limited, and more soil moisture data sets were not used to verify the accuracy of the filled CCI soil moisture data. However, in some special areas, the state of soil moisture may be special, which may lead to uncertainty in data quality, so it is crucial to refine the filling work. The freeze/thaw state of near-surface soil characterizes the dormancy and activity of land surface processes (Zhao et al., 2011; Wang et al., 2019; Hu et al., 2019). Since frozen soil cannot be used to retrieve soil moisture, we selected the 2002-2019 Global AMSR-E/2 Near-surface Freeze/Thaw state (0.25°) dataset to mask the frozen water part (Tianjie, 2018). The daily data is synthesized monthly. During the synthesis process, as long as there is liquid water in the month, the pixel is saved, and mask data is generated. In the remaining periods (1981.07-2002.05, 2020.01-2022.12), we performed monthly fusion on the valid mask data within the effective period, obtained mask data for each month, and applied these masks to the filled data.

https://doi.org/10.5194/essd-2024-200
Preprint. Discussion started: 15 July 2024

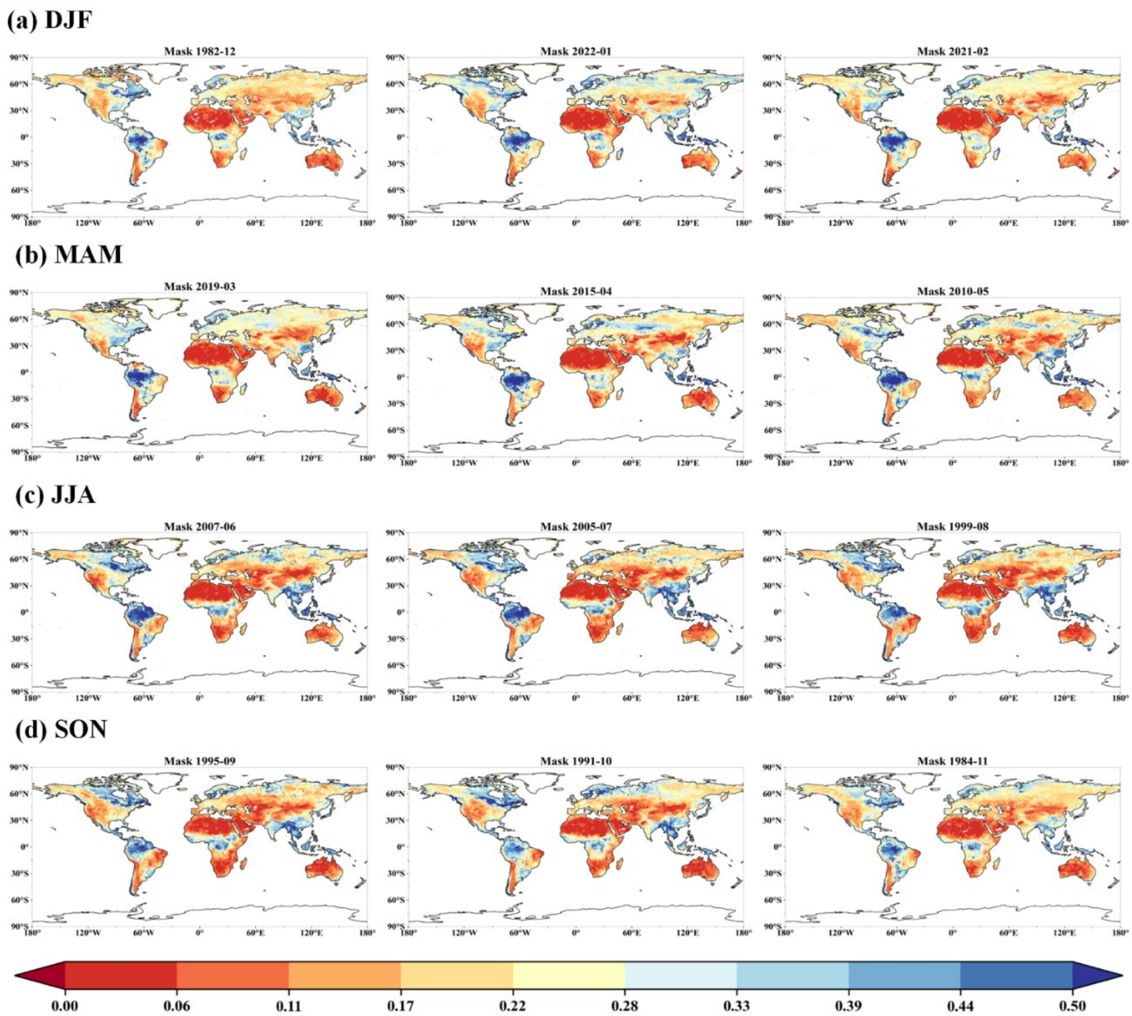

Figure 11: GSSM dataset after masking out frozen water.

5. Data availability

The global seamless soil moisture dataset from 1981 to 2022 dataset GSSM is available from https://data.tpdc.ac.cn/en/disallow/0f28a9b5-92eb-470a-80fe-472aa50a136f (last access: April 26, 2024) (Sun Hao, 2024).

6. Conclusions

The main contributions of this article are mainly reflected in three aspects: First, we propose a fitting beta CDF matching method that is more consistent with soil moisture data, while taking into account the boundary value selection problem in the matching process, which can ensure the characteristics of the soil moisture time series; Second, we used machine learning

XGBoost model to fill in the corrected data to solve the problem of low spatial coverage of soil moisture products. Finally, the dataset was obtained, namely long-term seamless CCI/SMAP monthly data soil moisture products (GSSM). By obtaining this dataset, researchers can take into account the advantages of long time range, and high spatial coverage soil moisture products.

Acknowledgements.

We sincerely thank the ISMN organisation for supplying in situ data. This research is supported by the Beijing Municipal Natural Science Foundation under Grant 6222045, and the China Fundamental Research Funds for Central University under Grant 2024JCCXDC03.

Competing interests.

The contact author has declared that none of the authors has any competing interests.

References

Abourizk, S., Halpin, D., and Wilson, J.: Fitting Beta Distributions Based on Sample Data, Journal of Construction Engineering and Management-asce - J CONSTR ENG MANAGE-ASCE, 120, 10.1061/(ASCE)0733-9364(1994)120:2(288), 1994.

Afshar, M. H. and Yilmaz, M. T.: The added utility of nonlinear methods compared to linear methods in rescaling soil moisture products, Remote Sens. Environ., 196, 224-237, 10.1016/j.rse.2017.05.017, 2017.

Aires, F., Weston, P., de Rosnay, P., and Fairbairn, D.: Statistical approaches to assimilate ASCAT soil moisture information-I. Methodologies and first assessment, Q. J. R. Meteorol. Soc., 147, 1823-1852, 10.1002/qj.3997, 2021.

Al-Yaari, A., Wigneron, J. P., Kerr, Y., Rodriguez-Fernandez, N., O'Neill, P. E., Jackson, T. J., De Lannoy, G. J. M., Al Bitar, A., Mialon, A., Richaume, P., Walker, J. P., Mahmoodi, A., and Yueh, S.: Evaluating soil moisture retrievals from ESA's SMOS and NASA's SMAP brightness temperature datasets, Remote Sens. Environ., 193, 257-273, 10.1016/j.rse.2017.03.010, 2017.

Anderson, M. C., Norman, J. M., Mecikalski, J. R., Otkin, J. A., and Kustas, W. P.: A climatological study of evapotranspiration and moisture stress across the continental United States based on thermal remote sensing: 2. Surface moisture climatology, J. Geophys. Res.-Atmos., 112, 13, 10.1029/2006jd007507, 2007.

Babaeian, E., Sadeghi, M., Jones, S. B., Montzka, C., Vereecken, H., and Tuller, M.: Ground, Proximal, and Satellite Remote Sensing of Soil Moisture, Rev. Geophys., 57, 530-616, 10.1029/2018rg000618, 2019.

Bai, L., Lv, X., and Li, X. J.: Evaluation of Two SMAP Soil Moisture Retrievals Using Modeled- and Ground-Based Measurements, Remote Sens., 11, 19, 10.3390/rs11242891, 2019.

Bertoldi, G., Claudia, N., Brenner, J., Castelli, M., Greifeneder, F., Niedrist, G., Seeber, J., and Tappeiner, U.: The role of soil moisture on the coevolution of soil and vegetation in mountain grasslands,

Bi, H. Y., Ma, J. W., Zheng, W. J., and Zeng, J. Y.: Comparison of soil moisture in GLDAS model simulations and in situ observations over the Tibetan Plateau, J. Geophys. Res.-Atmos., 121, 2658-2678, 10.1002/2015jd024131, 2016.

Brocca, L., Melone, F., Moramarco, T., and Morbidelli, R.: Soil moisture temporal stability over experimental areas in Central Italy, Geoderma, 148, 364-374, 10.1016/j.geoderma.2008.11.004, 2009.

Brocca, L., Hasenauer, S., Lacava, T., Melone, F., Moramarco, T., Wagner, W., Dorigo, W., Matgen, P., Martínez-Fernández, J., Llorens, P., Latron, J., Martin, C., and Bittelli, M.: Soil moisture estimation through ASCAT and AMSR-E sensors: An intercomparison and validation study across Europe, Remote Sens. Environ., 115, 3390-3408, 10.1016/j.rse.2011.08.003, 2011.

https://doi.org/10.5194/essd-2024-200
Preprint. Discussion started: 15 July 2024

Chan, S. K., Bindlish, R., O'Neill, P. E., Njoku, E., Jackson, T., Colliander, A., Chen, F., Burgin, M., Dunbar, S., Piepmeier, J., Yueh, S., Entekhabi, D., Cosh, M. H., Caldwell, T., Walker, J., Wu, X. L., Berg, A., Rowlandson, T., Pacheco, A., McNairn, H., Thibeault, M., Martínez-Fernández, J., González-Zamora, A., Seyfried, M., Bosch, D., Starks, P., Goodrich, D., Prueger, J., Palecki, M., Small, E. E., Zreda, M., Calvet, J. C., Crow, W. T., and Kerr, Y.: Assessment of the SMAP Passive Soil Moisture Product, IEEE Trans. Geosci. Remote Sensing, 54, 4994-5007, 10.1109/tgrs.2016.2561938, 2016.

Chen, J., Shi, X. Y., Gu, L., Wu, G. Y., Su, T. H., Wang, H. M., Kim, J. S., Zhang, L. P., and Xiong, L. H.: Impacts of climate warming on global floods and their implication to current flood defense standards, J. Hydrol., 618, 15, 10.1016/j.jhydrol.2023.129236, 2023.

Colliander, A., Jackson, T. J., Bindlish, R., Chan, S., Das, N., Kim, S. B., Cosh, M. H., Dunbar, R. S., Dang, L., Pashaian, L., Asanuma, J., Aida, K., Berg, A., Rowlandson, T., Bosch, D., Caldwell, T., Caylor, K., Goodrich, D., al Jassar, H., Lopez-Baeza, E., Martínez-Fernández, J., González-Zamora, A., Livingston, S., McNairn, H., Pacheco, A., Moghaddam, M., Montzka, C., Notarnicola, C., Niedrist, G., Pellarin, T., Prueger, J., Pulliainen, J., Rautiainen, K., Ramos, J., Seyfried, M., Starks, P., Su, Z., Zeng, Y., van der Velde, R., Thibeault, M., Dorigo, W., Vreugdenhil, M., Walker, J. P., Wu, X., Monerris, A., O'Neill, P. E., Entekhabi, D., Njoku, E. G., and Yueh, S.: Validation of SMAP surface soil moisture products with core validation sites, Remote Sens. Environ., 191, 215-231, 10.1016/j.rse.2017.01.021, 2017.

Crow, W. T. and van den Berg, M. J.: An improved approach for estimating observation and model error parameters in soil moisture data assimilation, Water Resour. Res., 46, 12, 10.1029/2010wr009402, 2010.

Cui, C. Y., Xu, J., Zeng, J. Y., Chen, K. S., Bai, X. J., Lu, H., Chen, Q., and Zhao, T. J.: Soil Moisture Mapping from Satellites: An Intercomparison of SMAP, SMOS, FY3B, AMSR2, and ESA CCI over Two Dense Network Regions at Different Spatial Scales, Remote Sens., 10, 19, 10.3390/rs10010033, 2018.

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

Dorigo, W. A., Xaver, A., Vreugdenhil, M., Gruber, A., Hegyiová, A., Sanchis-Dufau, A. D., Zamojski, D., Cordes, C., Wagner, W., and Drusch, M.: Global Automated Quality Control of In Situ Soil Moisture Data from the International Soil Moisture Network, Vadose Zone J., 12, 21, 10.2136/vzj2012.0097, 2013.

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

Draper, C., Mahfouf, J. F., Calvet, J. C., Martin, E., and Wagner, W.: Assimilation of ASCAT near- surface soil moisture into the SIM hydrological model over France, Hydrol. Earth Syst. Sci., 15, 3829-3841, 10.5194/hess-15-3829-2011, 2011.

Draper, C. S., Walker, J. P., Steinle, P. J., de Jeu, R. A. M., and Holmes, T. R. H.: An evaluation of AMSR-E derived soil moisture over Australia, Remote Sens. Environ., 113, 703-710, 10.1016/j.rse.2008.11.011, 2009.

Drusch, M., Wood, E. F., and Gao, H.: Observation operators for the direct assimilation of TRMM microwave imager retrieved soil moisture, Geophys. Res. Lett., 32, 4, 10.1029/2005gl023623, 2005.

Entekhabi, D., Njoku, E. G., O'Neill, P. E., Kellogg, K. H., Crow, W. T., Edelstein, W. N., Entin, J. K., Goodman, S. D., Jackson, T. J., Johnson, J., Kimball, J., Piepmeier, J. R., Koster, R. D., Martin, N., McDonald, K. C., Moghaddam, M., Moran, S., Reichle, R., Shi, J. C., Spencer, M. W., Thurman, S. W., Tsang, L., and Van Zyl, J.: The Soil Moisture Active Passive (SMAP) Mission, Proc. IEEE, 98, 704-716, 10.1109/jproc.2010.2043918, 2010.

Escorihuela, M. J. and Quintana-Seguí, P.: Comparison of remote sensing and simulated soil moisture datasets in Mediterranean landscapes, Remote Sens. Environ., 180, 99-114, https://doi.org/10.1016/j.rse.2016.02.046, 2016.

Ford, T. W. and Quiring, S. M.: Comparison of Contemporary In Situ, Model, and Satellite Remote Sensing Soil Moisture With a Focus on Drought Monitoring, Water Resour. Res., 55, 1565-1582, 10.1029/2018wr024039, 2019.

Gianotti, D. J. S., Salvucci, G. D., Akbar, R., McColl, K. A., Cuenca, R., and Entekhabi, D.: Landscape Water Storage and Subsurface Correlation From Satellite Surface Soil Moisture and Precipitation Observations, Water Resour. Res., 55, 9111-9132, 10.1029/2019wr025332, 2019.

https://doi.org/10.5194/essd-2024-200
Preprint. Discussion started: 15 July 2024

González-Zamora, A., Sánchez, N., Pablos, M., and Martínez-Fernández, J.: CCI soil moisture assessment with SMOS soil moisture and <i>in situ</i> data under different environmental conditions and spatial scales in Spain, Remote Sens. Environ., 225, 469-482, 10.1016/j.rse.2018.02.010, 2019.

Green, J. K., Seneviratne, S. I., Berg, A. M., Findell, K. L., Hagemann, S., Lawrence, D. M., and Gentine, P.: Large influence of soil moisture on long-term terrestrial carbon uptake, Nature, 565, 476-479, 10.1038/s41586-018-0848-x, 2019.

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

Hu, T. X., Zhao, T. J., Zhao, K. G., and Shi, J. C.: A continuous global record of near-surface soil freeze/thaw status from AMSR-E and AMSR2 data, Int. J. Remote Sens., 40, 6993-7016, 10.1080/01431161.2019.1597307, 2019.

Ji, X., Li, Y., Luo, X., He, D., Guo, R., Wang, J., Bai, Y., Yue, C., and Liu, C.: Evaluation of bias correction methods for APHRODITE data to improve hydrologic simulation in a large Himalayan basin, Atmospheric Research, 242, 104964, https://doi.org/10.1016/j.atmosres.2020.104964, 2020.

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., and Kumar, V.: Machine Learning for the Geosciences: Challenges and Opportunities, IEEE Trans. Knowl. Data Eng., 31, 1544-1554, 10.1109/tkde.2018.2861006, 2019.

Karthikeyan, L., Pan, M., Wanders, N., Kumar, D. N., and Wood, E. F.: Four decades of microwave satellite soil moisture observations: Part 1. A review of retrieval algorithms, Adv. Water Resour., 109, 106-120, 10.1016/j.advwatres.2017.09.006, 2017.

Kim, H., Parinussa, R., Konings, A. G., Wagner, W., Cosh, M. H., Lakshmi, V., Zohaib, M., and Choi, M.: Global-scale assessment and combination of SMAP with ASCAT (active) and AMSR2 (passive) soil moisture products, Remote Sens. Environ., 204, 260-275, 10.1016/j.rse.2017.10.026, 2018.

Konings, A. G., Entekhabi, D., Chan, S. K., and Njoku, E. G.: Effect of Radiative Transfer Uncertainty on L-Band Radiometric Soil Moisture Retrieval, IEEE Trans. Geosci. Remote Sensing, 49, 2686-2698, 10.1109/tgrs.2011.2105495, 2011.

Kornelsen, K. C. and Coulibaly, P.: Reducing multiplicative bias of satellite soil moisture retrievals, Remote Sens. Environ., 165, 109-122, 10.1016/j.rse.2015.04.031, 2015.

Kumar, S. V., Dirmeyer, P. A., Peters-Lidard, C. D., Bindlish, R., and Bolten, J.: Information theoretic evaluation of satellite soil moisture retrievals, Remote Sens. Environ., 204, 392-400, 10.1016/j.rse.2017.10.016, 2018.

Lee, J. H., Zhao, C. F., and Kerr, Y.: Stochastic Bias Correction and Uncertainty Estimation of Satellite-Retrieved Soil Moisture Products, Remote Sens., 9, 10.3390/rs9080847, 2017.

Liu, Y. X., Yang, Y. P., and Song, J.: Variations in Global Soil Moisture During the Past Decades: Climate or Human Causes?, Water Resour. Res., 59, 23, 10.1029/2023wr034915, 2023.

Liu, Y. Y., Parinussa, R. M., Dorigo, W. A., De Jeu, R. A. M., Wagner, W., van Dijk, A., McCabe, M. F., and Evans, J. P.: Developing an improved soil moisture dataset by blending passive and active microwave satellite-based retrievals, Hydrol. Earth Syst. Sci., 15, 425-436, 10.5194/hess-15-425-2011, 2011.

Llamas, R. M., Guevara, M., Rorabaugh, D., Taufer, M., and Vargas, R.: Spatial Gap-Filling of ESA CCI Satellite-Derived Soil Moisture Based on Geostatistical Techniques and Multiple Regression, Remote Sens., 12, 22, 10.3390/rs12040665, 2020.

Ma, H. L., Zeng, J. Y., Chen, N. C., Zhang, X., Cosh, M. H., and Wang, W.: Satellite surface soil moisture from SMAP, SMOS, AMSR2 and ESA CCI: A comprehensive assessment using global ground-based observations, Remote Sens. Environ., 231, 14, 10.1016/j.rse.2019.111215, 2019.

Ma, H. L., Zeng, J. Y., Chen, N. C., Zhang, X., Li, X. J., Wigneron, J. P., and Ieee: COULD L-BAND SOIL MOISTURE PRODUCTS CAPTURE THE SOIL MOISTURE CLIMATOLOGY VARIATIONS IN TROPICAL RAINFORESTS?, IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Pasadena, CA, Jul 16-21, WOS:001098971603091, 3153-3156, 10.1109/igarss52108.2023.10282045, 2023.

Madelon, R., Rodriguez-Fernández, N. J., van der Schalie, R., Scanlon, T., Al Bitar, A., Kerr, Y. H., de Jeu, R., and Dorigo, W.: Toward the Removal of Model Dependency in Soil Moisture Climate Data Records by Using an <i>L</i>-Band Scaling Reference, IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens., 15, 831-848, 10.1109/jstars.2021.3137008, 2022.

Moesinger, L., Dorigo, W., de Jeu, R., van der Schalie, R., Scanlon, T., Teubner, I., and Forkel, M.: The global long-term microwave Vegetation Optical Depth Climate Archive (VODCA), Earth Syst. Sci. Data, 12, 177-196, 10.5194/essd-12-177-2020, 2020.

https://doi.org/10.5194/essd-2024-200
Preprint. Discussion started: 15 July 2024

Pan, Y., Zhu, Y. H., Lue, H. S., Yagci, A. L., Fu, X. L., Liu, E., Xu, H. T., Ding, Z. Z., and Liu, R. Y.: Accuracy of agricultural drought indices and analysis of agricultural drought characteristics in China between 2000 and 2019, Agric. Water Manage., 283, 14, 10.1016/j.agwat.2023.108305, 2023.

Preimesberger, W., Scanlon, T., Su, C. H., Gruber, A., and Dorigo, W.: Homogenization of Structural Breaks in the Global ESA CCI Soil Moisture Multisatellite Climate Data Record, IEEE Trans. Geosci. Remote Sensing, 59, 2845-2862, 10.1109/tgrs.2020.3012896, 2021.

Rahimzadeh-Bajgiran, P., Berg, A. A., Champagne, C., and Omasa, K.: Estimation of soil moisture using optical/thermal infrared remote sensing in the Canadian Prairies, ISPRS-J. Photogramm. Remote Sens., 83, 94-103, 10.1016/j.isprsjprs.2013.06.004, 2013.

Reichle, R. H. and Koster, R. D.: Bias reduction in short records of satellite soil moisture, Geophys. Res. Lett., 31, 4, 10.1029/2004gl020938, 2004.

Robinson, D. A., Campbell, C. S., Hopmans, J. W., Hornbuckle, B. K., Jones, S. B., Knight, R., Ogden, F., Selker, J., and Wendroth, O.: Soil moisture measurement for ecological and hydrological watershed-scale observatories: A review, Vadose Zone J., 7, 358-389, 10.2136/vzj2007.0143, 2008.

Sadri, S., Wood, E. F., and Pan, M.: Developing a drought-monitoring index for the contiguous US using SMAP, Hydrol. Earth Syst. Sci., 22, 6611-6626, 10.5194/hess-22-6611-2018, 2018.

Sadri, S., Pan, M., Wada, Y., Vergopolan, N., Sheffield, J., Famiglietti, J. S., Kerr, Y., and Wood, E.: A global near-real-time soil moisture index monitor for food security using integrated SMOS and SMAP, Remote Sens. Environ., 246, 22, 10.1016/j.rse.2020.111864, 2020.

Sheffield, J. and Wood, E. F.: Global trends and variability in soil moisture and drought characteristics, 1950-2000, from observation-driven Simulations of the terrestrial hydrologic cycle, J. Clim., 21, 432-458, 10.1175/2007jcli1822.1, 2008.

Sheffield, J., Goteti, G., Wen, F. H., and Wood, E. F.: A simulated soil moisture based drought analysis for the United States, J. Geophys. Res.-Atmos., 109, 19, 10.1029/2004jd005182, 2004.

Shellito, P. J., Small, E. E., and Cosh, M. H.: Calibration of Noah Soil Hydraulic Property Parameters Using Surface Soil Moisture from SMOS and Basinwide In Situ Observations, Journal of Hydrometeorology, 17, 2275-2292, https://doi.org/10.1175/JHM-D-15-0153.1, 2016.

Su, C. H., Ryu, D., Western, A. W., and Wagner, W.: De-noising of passive and active microwave satellite soil moisture time series, Geophys. Res. Lett., 40, 3624-3630, 10.1002/grl.50695, 2013.

Sun, H. and Cui, Y. J.: Evaluating Downscaling Factors of Microwave Satellite Soil Moisture Based on Machine Learning Method, Remote Sens., 13, 16, 10.3390/rs13010133, 2021.

Sun, H. and Xu, Q.: Evaluating Machine Learning and Geostatistical Methods for Spatial Gap-Filling of Monthly ESA CCI Soil Moisture in China, Remote Sens., 13, 18, 10.3390/rs13142848, 2021.

Sun, H., Xu, Q., Wang, Y. J., Zhao, Z. Y., Zhang, X. H., Liu, H., and Gao, J. H.: Agricultural drought dynamics in China during 1982-2020: a depiction with satellite remotely sensed soil moisture, GISci. Remote Sens., 60, 28, 10.1080/15481603.2023.2257469, 2023.

Sun Hao, W. Y.: GSSM: A global long term seamless soil moisture dataset (1981-2022), National Tibetan Plateau Data Center [dataset], 10.11888/Terre.tpdc.301189, 2024.

Tianjie, Z.: 2002-2019 Global AMSR-E/2 Near-surface Freeze/Thaw state (0.25°), A Big Earth Data Platform for Three Poles [dataset], 10.11888/Glacio.tpdc.270890, 2018.

Vereecken, H., Huisman, J. A., Bogena, H., Vanderborght, J., Vrugt, J. A., and Hopmans, J. W.: On the value of soil moisture measurements in vadose zone hydrology: A review, Water Resour. Res., 44, 21, 10.1029/2008wr006829, 2008.

Wang, H. P., Song, J. Q., Zhao, C. W., Yang, X. R., Leng, H. Z., and Zhou, N.: Validation of the multi-satellite merged sea surface salinity in the South China Sea, Journal of Oceanology and Limnology, 41, 2033-2044, 10.1007/s00343-022-2187-x, 2023.

Wang, P. K., Zhao, T. J., Shi, J. C., Hu, T. X., Roy, A., Qiu, Y. B., and Lu, H.: Parameterization of the freeze/thaw discriminant function algorithm using dense <i>in-situ</i> observation network data, Int. J. Digit. Earth, 12, 980-994, 10.1080/17538947.2018.1452300, 2019.

Yang, H. X., Wang, Q. M., Zhao, W., and Atkinson, P. M.: Reconstruction of Historical SMAP Soil Moisture Dataset From 1979 to 2015 Using CCI Time-Series, IEEE Trans. Geosci. Remote Sensing, 62, 19, 10.1109/tgrs.2024.3360092, 2024.

https://doi.org/10.5194/essd-2024-200
Preprint. Discussion started: 15 July 2024

Yao, P. P., Lu, H., Shi, J. C., Zhao, T. J., Yang, K., Cosh, M. H., Gianotti, D. J. S., and Entekhabi, D.: A long term global daily soil moisture dataset derived from AMSR-E and AMSR2 (2002-2019), Sci. Data, 8, 16, 10.1038/s41597-021-00925-8, 2021.

Yao, P. P., Lu, H., Zhao, T. J., Wu, S. L., Peng, Z. Q., Cosh, M. H., Jia, L., Yang, K., Zhang, P., and Shi, J. C.: A global daily soil moisture dataset derived from Chinese FengYun Microwave Radiation Imager (MWRI)(2010-2019), Sci. Data, 10, 16, 10.1038/s41597-023-02007-3, 2023.

Zhang, P., Yu, H. B., Gao, Y. B., and Zhang, Q. F.: Evaluation of Remote Sensing and Reanalysis Products for Global Soil Moisture Characteristics, Sustainability, 15, 27, 10.3390/su15119112, 2023.

Zhang, Q., Yuan, Q. Q., Jin, T. Y., Song, M. P., and Sun, F. J.: SGD-SM 2.0: an improved seamless global daily soil moisture long-term dataset from 2002 to 2022, Earth Syst. Sci. Data, 14, 4473-4488, 10.5194/essd-14-4473-2022, 2022.

Zhao, T. J., Zhang, L. X., Jiang, L. M., Zhao, S. J., Chai, L. N., and Jin, R.: A new soil freeze/thaw discriminant algorithm using AMSR-E passive microwave imagery, Hydrol. Process., 25, 1704-1716, 10.1002/hyp.7930, 2011.