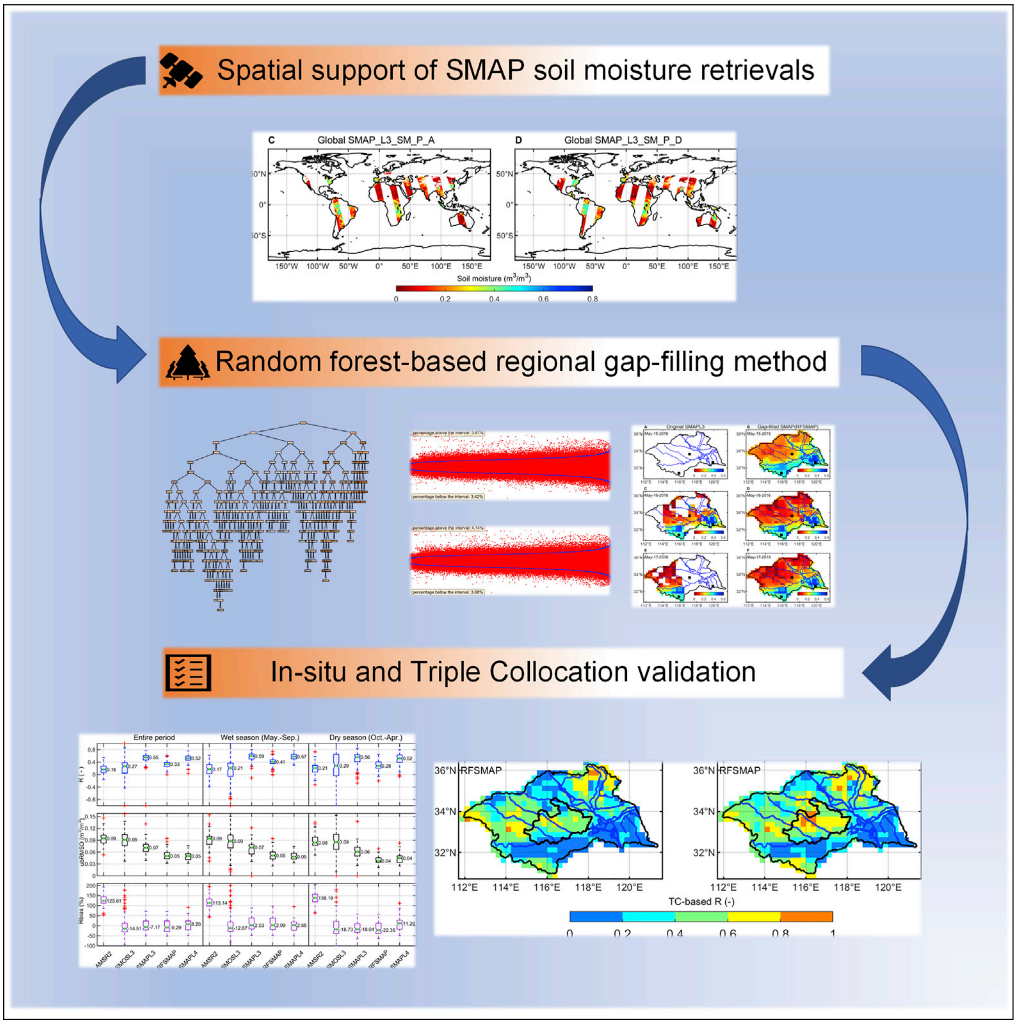
iScience

CellPress
OPEN ACCESS

Article

A reduced latency regional gap-filling method for SMAP using random forest regression



Spatial support of SMAP soil moisture retrievals

Random forest-based regional gap-filling method

In-situ and Triple Collocation validation

Xiaoyi Wang, Haishen Lü, Wade T. Crow, ..., Jianbin Su, Jingyao Zheng, Qiqi Gou

lvhaishen@hhu.edu.cn

Highlights

A gap-filling model boosts fast data availability of SMAP L3 soil moisture

Ground and triple collocation validation show consistent results for generated data

The seamless gap-filled data by integrating SMAP L3 form a unique source

Wang et al., iScience *26*, 105853
January 20, 2023 © 2022 The Author(s).
https://doi.org/10.1016/j.isci.2022.105853

iScience

CellPress OPEN ACCESS

Article

A reduced latency regional gap-filling method for SMAP using random forest regression

Xiaoyi Wang,[1,2] Haishen Lü,[1,2,6,*] Wade T. Crow,[3] Gerald Corzo,[4] Yonghua Zhu,[1,2] Jianbin Su,[5] Jingyao Zheng,[1,2] and Qiqi Gou[1,2]

SUMMARY

The soil moisture active/passive (SMAP) mission represents a significant advance in measuring soil moisture from satellites. However, its large spatial-temporal data gaps limit the use of its values in near-real-time (NRT) applications. Considering this, the study uses NRT operational metadata (precipitation and skin temperature), together with some surface parameterization information, to feed into a random forest model to retrieve the missing values of the SMAP L3 soil moisture product. This practice was tested in filling the missing points for both SMAP descending (6:00 AM) and ascending orbits (6:00 PM) in a crop-dominated area from 2015 to 2019. The trained models with optimized hyper-parameters show the goodness of fit ($R^2 \geq 0.86$), and their resulting gap-filled estimates were compared against a range of competing products with *in situ* and triple collocation validation. This gap-filling scheme driven by low-latency data sources is first attempted to enhance NRT spatiotemporal support for SMAP L3 soil moisture.

INTRODUCTION

Soil moisture observations are of great importance for hydro-meteorological and agricultural applications.[1–4] The growing recognition of the role of soil moisture underscores the need to obtain continuous, high-resolution quasi-global soil moisture data products in near real time (NRT) from space. During the past four decades, microwave-based satellite observations have proven to be an effective tool for satisfying this need.[5] During this period, a progressive series of experiments verified by truck-mounted sensors, aircraft, and space-borne sensors demonstrated that passive microwave radiometry can be applied to accurately retrieve surface soil moisture data (top ~5 cm).[6]

Soil moisture active/passive (SMAP) is the first mission designed to combine active and passive sensors to provide NRT soil moisture data and recognize frozen/thawed states on the land surface.[7] The volumetric accuracy goal for SMAP soil moisture retrievals is 0.04 m^3/m^3 (unbiased Root-Mean-Squared Error, ubRMSE) for the case of volumetric water vegetation content less than 5 kg/m^2. Extensive validation activity has demonstrated the high quality of SMAP soil moisture products,[8–11] as well as the potential feasibility of operational applications of SMAP-derived soil moisture products, e.g., flood modeling,[12] irrigation mapping,[13] and drought monitoring.[14]

Nevertheless, daily gaps often occur in level 2 and 3 SMAP products, which can limit their application. This issue is often more pressing in areas where soil moisture retrieval fails, or is flagged as unreliable, due to radio frequency interference (RFI), dense vegetation, or intense rainfall. To enhance the accuracy, vertical support, and spatiotemporal coverage of its level 2 and 3 soil moisture products, the SMAP mission also generates a time-continuous data assimilation product (SMAP L4) based on the assimilation of SMAP T_b measurements into a land surface model (LSM).[15] However, these advantages come at the expense of slightly increased data latency (average of about 2.5 days) due to a time lag incurred by the use of gauge-based precipitation as a required input for the SMAP L4 analysis.[16] In addition, because the SMAP L4 system is based on the assimilation of rescaled SMAP brightness temperature, the climatology of surface soil moisture estimates provided by the SMAP L4 analysis is not consistent with SMAP L2/L3 retrieval products.[16] Therefore, there remains the potential need for high-quality gap filling of SMAP L2/L3 products.

In addition, several past studies have made efforts on improving the temporal availability of satellite soil moisture products. The synergistic use of multiple Sun-synchronous orbit satellites can rapidly improve

[1] State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, National Cooperative Innovation Center for Water Safety and Hydro-science, College of Hydrology and Water Resources, Hohai University, Nanjing 210098, China

[2] Joint International Research Laboratory of Global Change and Water Cycle, Hohai University, Nanjing 210098, China

[3] USDA-ARS Hydrology and Remote Sensing Laboratory, Beltsville, MD 20705-2350, USA

[4] Hydroinformatics Chair Group, IHE Delft Institute for Water Education, 2611AX Delft, the Netherlands

[5] National Tibetan Plateau Data Center, Key Laboratory of Tibetan Environmental Changes and Land Surface Processes, Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing 100101, China

[6] Lead contact

*Correspondence: lvhaishen@hhu.edu.cn

https://doi.org/10.1016/j.isci.2022.105853

the spatial and temporal support of an individual satellite. Liu et al.[17] explored the possibility of using SMAP products directly for the gap-filling of the essential climate variable soil moisture in Europe. Kim and Lakshmi[18] applied Cyclone Global Navigation Satellite System–derived signal-to-noise ratio data in estimating soil moisture and filling them into the gap of missing spatial and temporal values in SMAP. Gruber et al.[19] used the triple collocation (TC) method for merging soil moisture retrievals from spaceborne active and passive microwave instruments based on weighted averaging of the error characteristics of individual data sets. However, such synergistic uses rely heavily on the availability of different satellite observations, and the mismatches of their overpass time could introduce errors into the ultimate integrated products. In addition, geostatistical techniques[20] and multiple regression have also been provided as alternatives for gap-filling,[21] where such interpolation methods fail when handling large spatial or temporal gaps.[22]

Machine learning (ML), as a powerful tool for processing nonlinear problems, can feasibly be used to capture correlations between satellite soil moisture (or its decomposition modes) and complementary information, mainly including meteorological forcing, geographic information, and vegetation conditions. Previous works have demonstrated that ML is a more robust way for filling gaps in satellite soil moisture retrievals than using geostatistics.[22–25] However, some predictors from satellites (e.g., Moderate Resolution Imaging Spectroradiometer Land Surface Temperature [LST] and Normalized Difference Vegetation Index) are often only applicable to clear-sky conditions due to cloud contamination.[26] In this way, the gaps would be ignored (removed) or prepopulated before running the ML model, which introduces uncertainty into the predicted results. In addition, the spatial/temporal transferability of the trained model is still vague (possible overfitting and underfitting) because the substantially large parameter space, especially hyper-parameter optimization and meta-modeling, is computationally expensive due to the need to train a large number of model configurations.[27] Therefore, how to efficiently perform a hyperparametric search is a current challenge in the field of ML.

Given the challenges discussed, the objectives of the present research are to (1) explore the suitability of NRT satellite precipitation and reanalysis data for SMAP L3 gap-filling, (2) report the utility of the successive halving search (SHS) in determining hyper-parameters for the random forest model, and (3) combine *in situ* validation and TC analysis to test the performance of gap-filling data sets and multiple on-orbit satellite products. The study area of this research is the Huai River basin of China (hereinafter referred to as the HRB), which is dominated by extensively irrigated cropland and has an urgent need for NRT soil moisture monitoring. The full names of the relevant specific organizations, algorithms and products can be viewed in Table 1.

Region of interest

Study area

The selected study area, the 270,000 km^2 HRB, is located between 30°55′—36°36′ N and 111°55′—121°25′ E (Figure 1). Because of its location in the transitional zone between the East Asian monsoon humid and the semi-humid regions, its weather is complex and highly variable with ample rainfall.[28] Its general meteorological and hydrographic characteristics are summarized by an average annual temperature of 11°C–16°C, average annual pan evaporation of 900–1500 mm, and average annual precipitation of 888 mm. The precipitation during the wet season (June to September) accounts for 50%–80% of the annual total precipitation. The region is prone to hydrologic extremes and has a historical record of frequent floods and drought. These events have important food security consequences because, as a major grain production base, the HRB has a very high area percentage of cropland[29] and provides around 17% of total grain yield in China together with a high population density of more than 160 million.[30] It is an area where soil moisture monitoring is critically important and, as a result, has been targeted for the development of ground-based hydrological monitoring networks.

Unfortunately, the HRB region is also characterized by poor temporal coverage in SMAP L3 retrievals (Figure 2), which, therefore, does not meet the requirements for reliable regional continuous monitoring of soil moisture. The HRB region is affected by aggressive flagging of SMAP L3 retrievals, likely linked to the presence of local anthropogenic RFI sources. In addition, it commonly experiences a number of natural characteristics (e.g., active rainfall, dense vegetation, and complex topography) that can cause failures in soil moisture retrieval algorithms.[29] As such it represents an excellent test-case location for examining the universality of latency-reduced gap-filling strategy over a cultivation-dominated area. In addition,

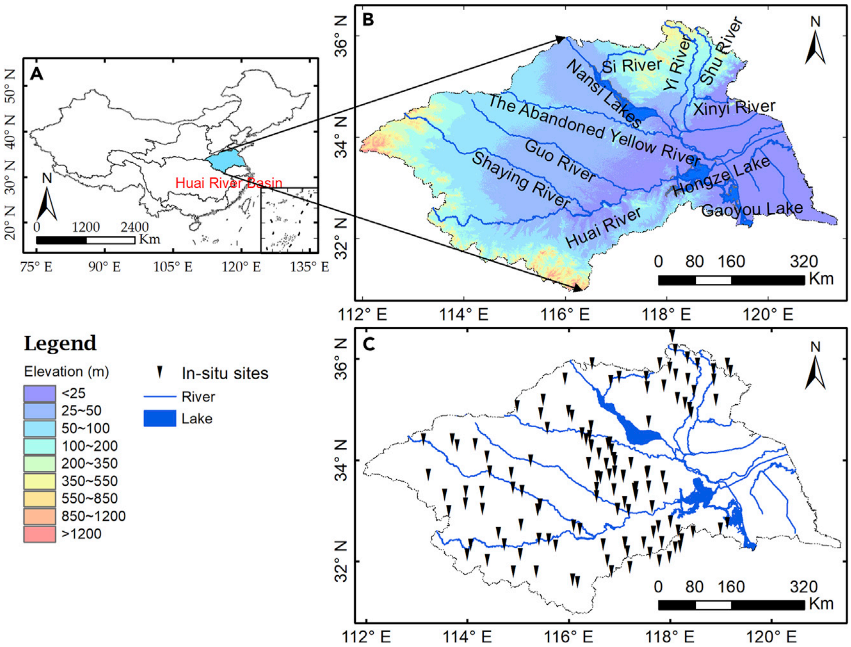


Figure 1. Basic information of the study area

(A) Location of the HRB in China.

(B) The DEM map of the basin with labeled river branches.

(C) The distribution of the HRB *in situ* soil moisture measurement sites (98) (sites labeled with gray inverted triangles).

the abundant availability of *in situ* observations by the Ministry of Water Resources of China (MWR) in the HRB (Figure 1C) allows for the reliable benchmarking of soil moisture data products within the region.

Data resources

Three diverse sources of data (Table 2) within the study area were used to achieve the study goals: satellite observations, LSM outputs, and *in situ* observations acquired from March 31, 2015, to December 31, 2019. Specifically, two passive microwave soil moisture products (Soil Moisture and Ocean Salinity [SMOS] and Advanced Microwave Scanning Radiometer 2 [AMSR2]) and one SMAP-derived LSM data set (SMAP L4) were applied to make intercomparisons with the gap-filled products. All satellite products are retrieved from both daily ascending and descending orbits, while LSM outputs are provided hourly or 3-hourly.

Gap-filling object product

The National Aeronautics and Space Administration SMAP mission started data acquisition on March 31, 2015, and aims to measure the amount of water in the surface soil and freeze/thaw state everywhere on Earth from space. Its observation system was originally based on a combined L-band radar (1.26 and 1.29 GHz) and radiometer (1.41 GHz) to leverage the relative advantages from both active and passive microwave remote sensing for surface soil moisture (i.e., top ~5 cm). However, on July 7, 2015, its radar stopped transmitting due to an anomaly involving its high-power amplifier. Nevertheless, SMAP's radiometer has continued to operate and gather scientific data regularly. Here, the SMAP level 3 soil moisture product (SMAP L3 for short in this study) for both ascending and descending SMAP orbits was downloaded and resampled from EASE-Grid 2.0 onto a 0.25° × 0.25° resolution geographic projection.

Ancillary data for gap-filling

Global precipitation measurement product

The IMERG V06 Early Run (IMERG-E) product is freely available from the Goddard Earth Sciences Data and Information Services Center. It is a quasi-real-time product with a temporal data latency of 4 hours.[31] The spatial resolution of IMERG-E is 0.1°, and multiple temporal resolutions are available (i.e., 30 min, 3 h,

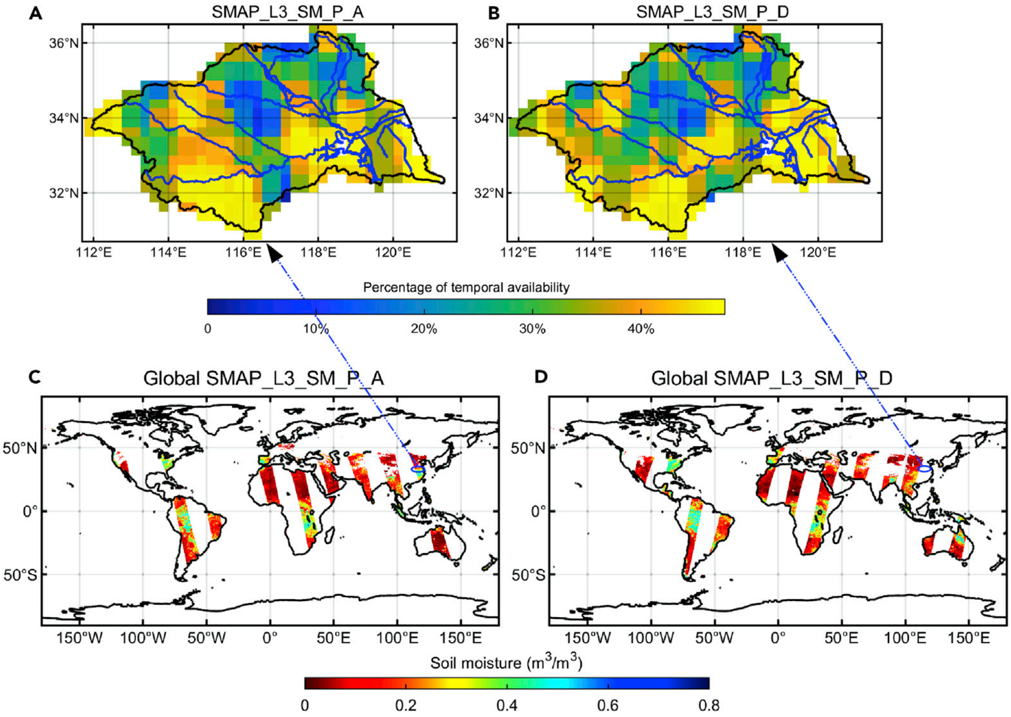


Figure 2. Percentage of data gaps in the SMAP soil moisture products

(A and B) The percentage of temporal data availability (i.e., days with retrievals divided by total days) of SMAP L3 over the HRB for ascending orbits (18:00 local solar time) and descending orbits (6:00 local solar time), where the blank-pixel area inside the basin is Hongze Lake.

(C and D) The global observation swath path of SMAP L3 on March 3, 2018, for ascending and descending orbits.

1 day, and 7 days following data acquisition).[32] IMERG-E possesses the capability for flood warning in South China,[33] and its 30-min product after preprocessing was used in this study as an input in the gap-filling model.

SMAP L1-L3 ancillary data

The Goddard Earth Observing System, Version 5 Forward Processing (GEOS-5 FP) atmospheric temperature[34] and IMERG data set are the only dynamic model products used in this study, and they have almost no gaps in space. In particular, the required SMAP 0~5 cm LST in this study was obtained by averaging the skin temperature and first-layer soil temperature products from the GEOS-5 FP system. Similarly, the corresponding SMAP ancillary static data files contain soil attribution (clay, sand, and bulk density), digital elevation model (DEM), slope, and roughness values.

Validation data

In situ soil moisture

The *in situ* soil moisture automonitoring system, operated by the MWR, provides long-term soil moisture records for three layers (i.e., 0–10 cm, 10–20 cm, and 20–40 cm beneath the ground surface). Depending on the site, measurements are recorded either daily or every 10 days at 8:00 local solar time. MWR applies quality-control processing to flag suspicious measurements[35] and rejects invalid sites. Within the HRB, a total of 98 sites (Figure 1C) have adequate sampling—defined as providing more than 100 records during the study period (March 31, 2015, to December 31, 2019). Because SMAP's effective detecting depth is usually shallow than 5 cm, only the first-layer (0–10 cm) *in situ* data set was used.

Advanced scatterometer soil moisture

Advanced scatterometer (ASCAT) is a real-aperture radar instrument that operates at the C-band (5.3 GHz, 5.7 cm wavelength) aboard the EUMETSAT MetOp-A (October 2006), MetOp-B (September 2012), and MetOp-C (November 2018) satellites.[36] It can retrieve soil moisture with a sensing depth of 2–5 cm.

Table 1. List of abbreviations

Abbreviation	Expansion
AMSR-E	Advanced Microwave Scanning Radiometer–Earth Observing System
AMSR2	Advanced Microwave Scanning Radiometer 2
ASCAT	Advanced Scatterometer
CATDS	Centre Aval de Traitement des Données
CLDAS	China Land Data Assimilation System v2.0
DEM	Digital Elevation Model
ECV-SM	The Essential Climate Variable soil moisture data sets
EUMETSAT	European Organisation for the Exploitation of Meteorological Satellites
HSAF	Hydrology and Water Management
IMERG	Integrated Multi-satellite Retrievals for GPM
JPL	NASA's Jet Propulsion Laboratory
GEOS-5 FP	Goddard Earth Observing System, Version 5, Forward Processing
GMAO	Global Modeling and Assimilation Office
GPM	Global Precipitation Measurement
GSFC	Goddard Space Flight Center
ISMN	International Soil Moisture Network
LST	Land Surface Temperature
MetOp	Meteorological Operational satellite program
MODIS	Moderate Resolution Imaging Spectroradiometer
MWR	Ministry of Water Resources of the People's Republic of China
NASA	National Aeronautics and Space Administration
NDVI	Normalized Difference Vegetation Index
NMIC-CMA	National Meteorological Information Center–China Meteorological Administration
NSIDC	National Snow and Ice Data Center
SMOS	Soil Moisture and Ocean Salinity
SMAP	Soil Moisture Active Passive
T_B	Brightness Temperature

ASCAT crosses the equator at 21:30 local solar time and 09:30 local solar time in descending and ascending orbits, respectively. This study used the composite ASCAT soil moisture product (hereafter referred to as ASCAT) derived at the Integrated Climate Data Center of Hamburg University[37] based on reprocessed version 5 of the EUMETSAT H-SAF H115 and H116 products. These composite data are sampled from 12.5 km swath orbit geometry to per grid cell (0.1° × 0.1°) and daily released for both ascending and descending paths. Composite ASCAT soil moisture retrievals are provided in relative units ranging between 0% (dry) and 100% (saturated). To acquire volumetric soil moisture, porosity data provided in each ASCAT composite file were extracted to obtain volumetric soil moisture estimates in m^3m^{-3} units.

SMOS soil moisture

As one of only two on-orbit satellites operating in the L-band (i.e., 1.43 GHz, 21 cm) frequency, European Space Agency's SMOS is the first satellite designed for measuring global surface soil moisture and ocean salinity.[38] SMOS has provided scientific soil moisture estimates twice a day (ascending/descending orbit, 6:00 AM/6:00 PM local solar time) on a 25-km EASE-Grid2 projection since November 2009.[39] The SMOS level 3 daily quasi-global soil moisture data set (SMOS L3) was publicly accessed from Centre Aval de Traitement des Données.

AMSR2 soil moisture

As the successor of the AMSR Earth Observing System sensor, AMSR2 was launched in May 2012 and designed for measuring microwave emissions from the surface and the atmosphere of the earth. It detects passive microwave frequencies from 6.925 to 89.3 GHz and began the release of scientific data on

Table 2. Overview of the products involved in the gap-filling model and validation

Sources (version)	Institution	Resolution support	Data latency (required)
Gap-filling data			
SMAP (SPL3SMP)	NASA-GSFC	36 km/d	50 h
Day of year (DOY)	/	0.25° × 0.25°	/
Latitude (LAT)			
Longitude (LON)			
Rainfall (IMERG Early Run)	NASA-GPM	0.1°/0.5 h	4 h
LST (GEOS-5 FP)	NASA-GMAO	0.25° × 0.3125°/h	7 h
DEM (DEM_M36_003)	NASA-JPL	36 km	/
Slope (DEMSLP_M36_00)			
Roughness (M36_002)			
Clay (M36_004)	NSIDC/NASA	36 km	/
Bulk (M36_004)			
Sand (M36_004)			
Validation data			
In situ data	MWR	/	Real-time
ASCAT (H-SAF V7)	EUMETSAT-HSAF-UHAM-ICDC	0.1° × 0.1°/d	12–36 h
SMOS (SMOS_L3)	European Space Agency	25 km/d	4 h
AMSR2 (6.9 GHz)	NASA	25 km/d	3 h
CLDAS (CLDAS v2.0)	NMIC-CMA	0.0625°/h	1 h
SMAP_L3 (SPL3SMP)	NASA	36km/d	50 h
SMAP_L4 (SPL4SMGP)	NASA	9km/3h	7 d

July 3, 2012.[40] Its gridded soil moisture product is based on a combination of daytime (10:30 PM local solar time) and nighttime (1:30 AM local solar time) data and covers more than 99% of the globe every 2 days.[41] This study used the Land Parameter Retrieval model soil moisture product based on 6.9-GHz AMSR2 observations.[42]

China Land Data Assimilation System v2.0 soil moisture product

An LSM output, the China Land Data Assimilation System v2.0 (CLDAS) soil moisture data set covering eastern Asia, was used for TC analysis in this study. The China Meteorological Data Service Center is in charge of the data distribution derived from CLDAS operation in real time (at a latency of 1 h) and in NRT (at a latency of 2 days) since 2008.[43] The CLDAS soil moisture products are averaged from a three-member ensemble of off-line LSMs (i.e., the Community Land Model version 3.5, the Common Land Model, and the Noah Multi-parameterization 1.4 Land Surface Model) driven by 40,000 automatic meteorological stations observations, satellite precipitation, and numerical weather predictions.[44] Compared with other LSM outputs, the CLDAS soil moisture product (hereinafter referred to as "CLDAS") shows a better performance with finer spatial-temporal resolution (0.0625° × 0.0625°, hourly). CLDAS is freely available from the National Science & Technology Infrastructure of China. In the present study, only CLDAS top-layer (0∼5 cm) soil moisture at 6:00 AM/PM was selected.

SMAP L4 soil moisture geophysical data

The SMAP L4 soil moisture geophysical data (SPL4SMGP) product[45] consists of 3-hourly soil moisture estimates obtained via the assimilation of SMAP L1C T_B observations into the GEOS-5 Catchment LSM,[46] with the inputs of the surface meteorological forcing data stream (including precipitation) from a global atmospheric model output by assimilating a very large number (greater than 10^7 per day) of conventional and satellite-based observations of the atmosphere.[47] It provides perfect spatial-temporal coverage in its soil moisture estimates by filling the gaps in SMAP observations due to orbit and land surface characteristics. However, this coverage comes at the expense of increased data latency. The forward processing of the Catchment model background takes nearly 3 days to complete, and the official data latency requirement for SPL4SMGP is within 7 days.

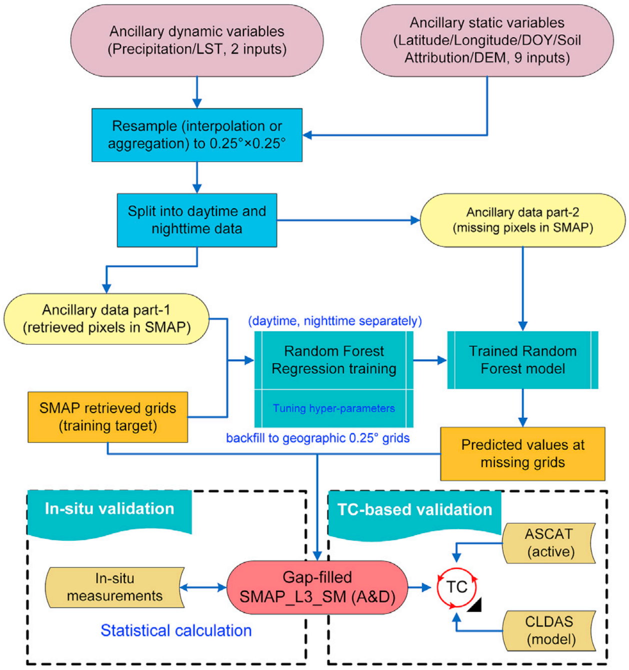

Ancillary dynamic variables (Precipitation/LST, 2 inputs)
Ancillary static variables (Latitude/Longitude/DOY/Soil Attribution/DEM, 9 inputs)
Resample (interpolation or aggregation) to 0.25°×0.25°
Split into daytime and nighttime data
Ancillary data part-2 (missing pixels in SMAP)
Ancillary data part-1 (retrieved pixels in SMAP)
(daytime, nighttime separately)
Random Forest Regression training
Tuning hyper-parameters
Trained Random Forest model
SMAP retrieved grids (training target)
backfill to geographic 0.25° grids
Predicted values at missing grids
In-situ validation
TC-based validation
ASCAT (active)
In-situ measurements
Gap-filled SMAP_L3_SM (A&D)
TC
CLDAS (model)
Statistical calculation

Figure 3. Work flowchart for gap-filling SMAP soil moisture products and validation of the subsequent continuous soil moisture product

Methodologies

This section describes (1) data preprocessing to a unified framework, (2) the construction of a random forest regression model for predicting values within missing pixels, and (3) the calculation of evaluation metrics for product validation. The entire work flowchart for the approach is shown in Figure 3.

Data preprocessing

To increase the frequency of data available for training the gap-filling model, this study omitted quality control procedures for the SMAP L3 product. However, quality control was applied to SMOS L3 retrievals before the intervalidation. Specifically, SMOS L3 pixels were rejected when their Data Quality Index was over 0.07 or equal to fill values that indicate retrievals failed. Similarly, pixels were also rejected if the $\mathrm{RFI}_{\mathrm{fraction}}$ (percentage of RFI) variable was higher than 0.3.[48]

Next, all inputs and training targets for the gap-filling model and participated evaluation products (ASCAT, AMSR2, SMOS_L3, SMAP_L4, and CLDAS) were resampled onto a fixed 0.25° grid by the nearest neighbor interpolation. In addition, all IMERG precipitation products were resampled to a daily time scale.

Random forest–based gap-filling approach

Random Forest (denoted by RF below) is a meta-estimator capable of performing both classification and regression tasks through the use of decision trees and their regressor.[49] RF fits many decision trees on various seed samples of the data set and uses averaging to improve its prediction accuracy and exert overfitting control. Similarly, its computational efficiency allows for quick applications to large data sets.[50]

Bagging, i.e., bootstrap aggregation sampling in the RF model, indicates training every decision tree regressor in the random forest on a different sample where sampling is done with replacement. Next, a collection of decision tree regressors that run in parallel without any mutual interaction determine the final prediction.

After data cleaning, a total of 742,851 sets of data (each set contains 11 features) participated in the model construction. In this article, the whole successful SMAP L3 retrievals were used as training data, which could be split into internal training and validation in each iteration by out-of-bag (oob) sampling after bagging. Finally, the predictor information located in these grids could drive the trained model to fill in the missing records.

Table 3. Hyper-parameters for both two models

Hyper-parameter	Initial range	Ascending model oob score(R^2) = 0.86	Descending model oob score (R^2) = 0.86
n_estimators	100–2000	1200	1400
max_depth	10–100	20	20
max_features	1/0.5/$\log_2$/sqrt	0.5	0.5
min_samples_leaf	2, 5, 9	8	5
min_samples_split	1, 2, 4	2	2

Hyper-parameter tuning and model training

The SHS is developed to overcome the inefficiency of the high-dimensionality hyper-parameter configuration space because the number of evaluations increases exponentially as the number of hyper-parameters increases.[51] The logic of the SHS is to assign a small amount of resources to the hyper-parameter combinations (assuming n sets) for evaluation, and at each iteration, half of the poorly performing hyper-parameter configurations are discarded, while the better performing half proceeds to the next iteration with a double budget until the final best hyper-parameter combination is determined.[52]

SHS-based algorithms as the built-in functions can be easily called from the scikit-learn library.[53] Accordingly, SHS-based random search (HalvingRandomSearchCV) and grid search (HalvingGridSearchCV) were adopted successively in this study to determine the hyper-parameters of the random forest model. The former is used to search a set of fuzzy values for certain hyper-parameters, and the latter is used to search the ultimate combination around fuzzy values. At the same time, the training process also sets 5-fold cross-validation to avoid overfitting. The initial hyper-parameter ranges and the ultimate hyper-parameter configuration are reported in Table 3.

Moreover, during the training stage, oob samples are automatically drawn from the training set for model reliability testing, whose performance can be measured by the coefficient of determination (i.e., oob score in the random forest model, referred to in Table 3). After obtaining the trained RF model, locations where SMAP L3 soil moisture was unsuccessful (or otherwise unavailable) were predicted and used to backfill the original SMAP L3 mesh network and produce the integrated, continuous data set.

Statistical metrics for direct comparison

The Pearson correlation coefficient (R), unbiased root mean square difference (ubRMSD), and relative bias (Rbias) are used to directly compare the gridded products against *in situ* soil moisture observations. The relative formulas for these three metrics are as follows:

$$R = \frac{\sum_{i=1}^{n}(OBS_i - \overline{OBS})(SAT_i - \overline{SAT})}{\sqrt{\sum_{i=1}^{n}(OBS_i - \overline{OBS})^2 \sum_{i=1}^{n}(SAT - \overline{SAT})^2}} \qquad \text{(Equation 1)}$$

$$Rbias = \frac{\sum_{i=1}^{n}(SAT_i - OBS_i)}{\sum_{i=1}^{n}(OBS_i)} \times 100\% \qquad \text{(Equation 2)}$$

$$ubRMSD = \sqrt{\frac{\sum_{i=1}^{n}(SAT_i - OBS_i)^2}{n} - (Rbias/100)^2} \qquad \text{(Equation 3)}$$

where n represents the total number of sampled retrievals; i is the node of time series; and OBS and SAT indicate *in situ* observations and the satellite-derived soil moisture, respectively. Accordingly, $\overline{OBS}$ and $\overline{SAT}$ express the mean value of *in situ* data and the satellite-derived soil moisture.

TC analysis

TC, proposed by Stoffelen,[54] is a common approach for estimating the random error variance of at least three collected, independent, measurement systems without access to a true error-free representation. The essential assumption of TC is the mutual independence of errors between the three select measurement systems.

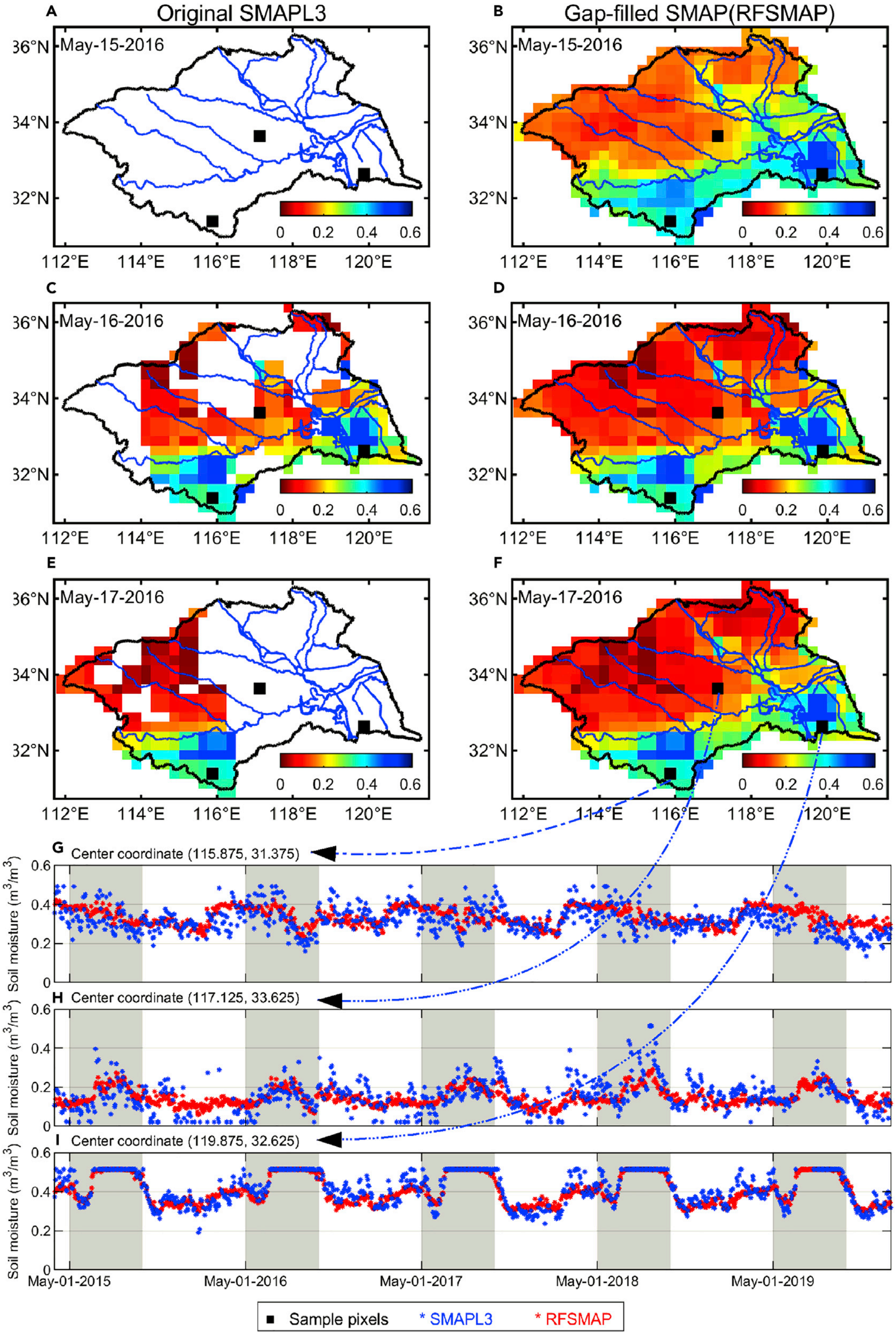

A
Original SMAPL3
May-15-2016
B
Gap-filled SMAP(RFSMAP)
May-15-2016
C
May-16-2016
D
May-16-2016
E
May-17-2016
F
May-17-2016
36°N
34°N
32°N
112°E
114°E
116°E
118°E
120°E
0
0.2
0.4
0.6
G
Center coordinate (115.875, 31.375)
H
Center coordinate (117.125, 33.625)
I
Center coordinate (119.875, 32.625)
Soil moisture (m^3/m^3)
May-01-2015
May-01-2016
May-01-2017
May-01-2018
May-01-2019
■ Sample pixels
* SMAPL3
* RFSMAP

Figure 4. Comparison of the gap-filling effects

(A–F) Comparison of original (A, C, and E) and gap-filled SMAP L3 (B, D, and F) retrievals for the SMAP descending path on three consecutive days (May 15, 2016–May 17, 2016).

(G–I) An example (multiyear) time series for three sample pixels labeled with black squares. The shadings represent the general flood season over the HRB.

For soil moisture, the TC triplet consists of an active microwave-based retrieval (*X*), a passive microwave-based retrieval (*Y*), and a model output (*Z*). This triplet is assumed to be linearly related to true soil moisture values via:

$$\begin{aligned} X &= \alpha_X + \beta_X T + \varepsilon_X \\ Y &= \alpha_Y + \beta_Y T + \varepsilon_Y \\ Z &= \alpha_Z + \beta_Z T + \varepsilon_Z \end{aligned} \qquad \text{(Equation 4)}$$

where α, β, and ε represent additive systematic errors, multiplicative systematic errors, and additive zero-mean random errors of each data set, respectively, compared with the true value *(T)*.

If the TC assumptions hold, then the error variances can be estimated via averaging the cross-multiplied differences among them.[55] The ubRMSD in each data set can be calculated based on the error variance below:

$$\begin{aligned} ubRMSD_X &= \sqrt{\sigma^2_{\varepsilon_X}} = \sqrt{\sigma^2_X - \frac{\sigma_{XY}\sigma_{XZ}}{\sigma_{YZ}}} \\ ubRMSD_Y &= \sqrt{\sigma^2_{\varepsilon_y}} = \sqrt{\sigma^2_Y - \frac{\sigma_{YX}\sigma_{YZ}}{\sigma_{XZ}}} \\ ubRMSD_Z &= \sqrt{\sigma^2_{\varepsilon_Z}} = \sqrt{\sigma^2_Z - \frac{\sigma_{XY}\sigma_{XZ}}{\sigma_{XY}}} \end{aligned} \qquad \text{(Equation 5)}$$

In particular, the extended TC method, proposed by McColl et al.,[56] was applied here to obtain the correlation coefficient estimates (R):

$$R_{i,T} = \sqrt{\frac{\beta_i^2 \sigma_T^2}{\beta_i^2 \sigma_T^2 + \sigma^2_{\varepsilon_i}}} \qquad \text{(Equation 6)}$$

where $R_{i,T}$ represents the correlation coefficient between each data set and the unknown truth of soil moisture. Here, the metrics for the TC-based ubRMSD in Equation 5 and the correlation coefficient ($R_{i,T}$) obtained in Equation 6 are used to evaluate various soil moisture products. Further details about TC can be found in the STAR Methods.

RESULTS

Figure 4 presents the example results on descending orbit for the comparison of the original SMAP L3 and integrated RF-SMAP for 3 days (May 15, 2016–May 17, 2016) and three time series of soil moisture at three different pixels in the HRB. The proposed method can describe the spatial distribution of soil moisture by filling the missing pixels in the original SMAP L3, but the specific data quality needs to be further verified and analyzed. In this section, the integrated data were separated into mutually exclusive "SMAP-retrieved" and "gap-filled" temporal periods data, namely SMAP L3 and RF-SMAP, to ensure that validation targets only soil moisture estimates generated by the RF-SMAP regression procedure (i.e., only gap-filled values), while the same validation procedure also been applied for the integrated data set (see Figures S4–S6).

Validation will examine the accuracy of RF-SMAP estimates relative to existing, independent soil moisture product (i.e., AMSR2, SMOS L3, and the SMAP DA product—SMAP L4) by (1) examining the temporal behavior of each product using observation records (*in situ* validation) and (2) examining their detailed spatial features (all but SMAP L4) pixel by pixel by applying TC (TC validation). In both cases, SMAP L3 and RF-SMAP retrievals will be evaluated relative to the AMSR2, SMOS L3, and SMAP L4 products.

In situ validation

Because the *in situ* daily acquisition is at local 8:00 AM, the satellite products acquired need to be selected on the daytime pass, e.g., AMSR2 descending product (1:30 PM), SMAP L3 descending product (6:00 AM), RF-SMAP descending product (6:00 AM), SMOS L3 ascending product (6:00 AM), and SMAP L4 (7:30 AM) analysis.

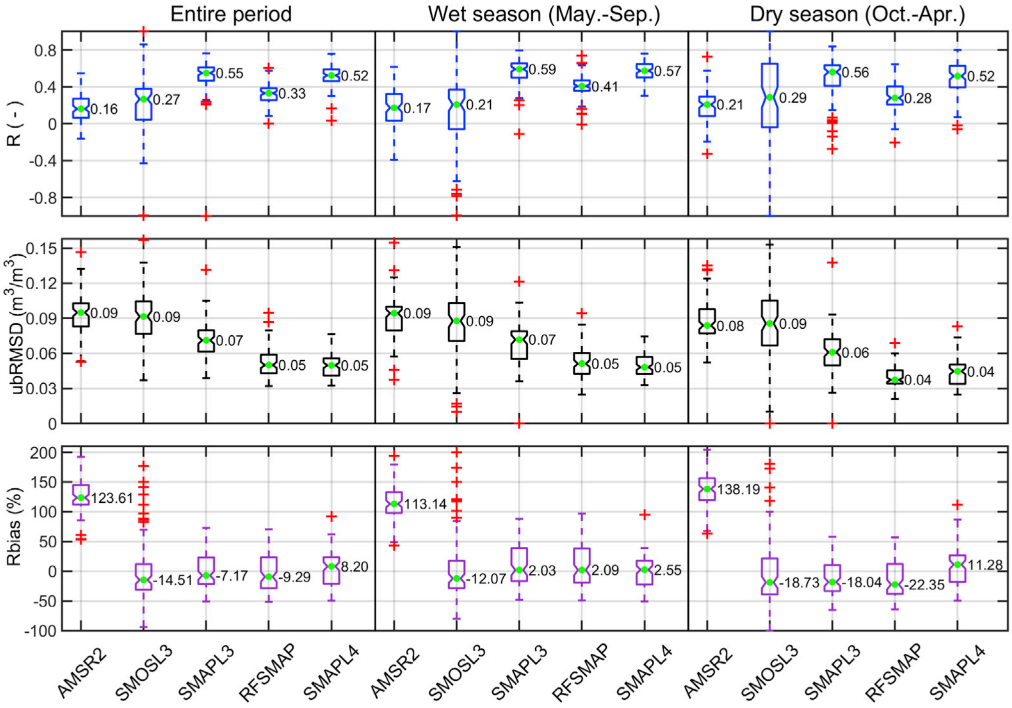


Figure 5. Boxplots for the spatial distribution of R, ubRMSD, and Rbias values for AMSR2, SMOS L3, SMAP L3, RF-SMAP, and SMAP L4 retrievals versus daytime *in situ* data

The left, middle, and right columns represent the metrics during the entire study period, the wet season, and the dry season, respectively.

Their performance during the whole study period, the wet season (May to September), and the dry season (October to April) was evaluated first. Boxplots (Figure 5) display the distribution of R, ubRMSD, and Rbias evaluation statistics calculated across all available *in situ* validation sites. SMAP L3 is, unexpectedly, slightly superior to SMAP L4—especially with regards to its median R values during all seasons.

After these two existing SMAP baseline products, in which the R values are above 0.5, the next best product is RF-SMAP. It has consistently good evaluation metrics with SMAP L3 and has slightly improved (median R = 0.41) during the wet season versus other periods. Furthermore, RF-SMAP holds the lowest median ubRMSD (≤0.05 m^3/m^3) among these intercompared products. However, a significant decline of retrieval quality during the dry season can be found in RF-SMAP—likely associated with SMAP's overall degraded retrieval capability (i.e., reduced number of collected granules) during the cold season (December to February).

Relatively poorer retrieval quality is observed in the AMSR2 and SMOS L3 products. AMSR2 illustrates the lowest R and overestimates *in situ* observations during all three time periods, and this tendency is greater during the dry season (median Rbias = 138.19%). SMOS L3 exhibits the most unstable performance—as reflected by the width of their performance boxplots in Figure 5, which may be related to known SMOS RFI issues in eastern Asia.[57] In addition, extreme overestimated outliers can be seen in SMOS L3 retrievals during both the wet and dry seasons.

Next, the temporal cumulative distribution function (CDF) tool was calculated for seasonal analysis (Figure 6). Overall, the shape of the CDF curves varies greatly among soil moisture data sets and between seasons. Nearly all AMSR2 retrievals are above the median values of *in situ* data for both the wet and dry seasons. The range of the CDF curves for AMSR2 and SMOS L3 is larger than that for the three SMAP-related data sets, indicating their larger dynamic range versus the *in situ* data. In contrast, the SMAP-related products show relatively good agreement with *in situ* data, and RF-SMAP retrievals largely inherit this good fit. A larger SMAP L3 discrepancy versus *in situ* data is seen during the wet season, which is likely due to errors being introduced in the vegetation correction during the retrieval.[58]

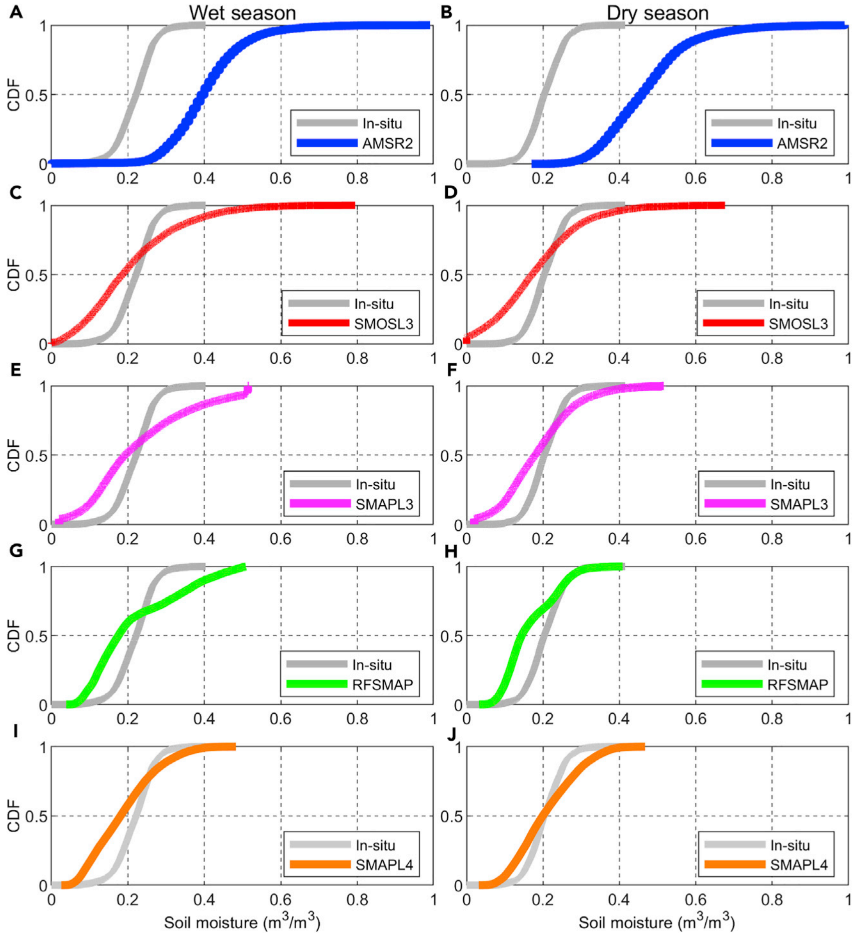


Figure 6. Comparison of the values distribution between the gap-filled data and in-situ data

(A–J) The CDF curves of AMSR2 (A and B, blue line), SMOS L3 (C and D, red line), SMAP L3 (E and F, carmine line), RF-SMAP for gaps (G and H, green line), and SMAP L4 (I and J, orange line) compared with *in situ* data (gray line) during the wet (left column) and dry (right column) seasons.

Despite the overall stable accuracy of RF-SMAP, there are significant geographic variations in its accuracy. Figure 7 shows *in situ*-based spatial metrics (the R, ubRMSD, and Rbias) sampled versus *in situ* soil moisture observations. Results suggest that SMAP L3 and SMAP L4 present the best correlation with ground reference observations across space, and RF-SMAP follows, while AMSR2 and SMOS L3 demonstrate the weakest correlations with significant bias. Besides, RF-SMAP can match the widespread good consistency of SMAP L3 and has a smaller error distribution although its R-values are degraded a little. This is explainable because RF-SMAP always predicts data under unfavorable retrieval conditions.

Beyond that, an interesting phenomenon is shown in the Rbias map: AMSR2 overestimates the *in situ* measurements while a widespread underestimation is seen in the other four data sets. This is consistent with previous studies suggesting that the Land Parameter Retrieval Model algorithm tends to overestimate soil moisture.[40] In addition, both SMAP L3 and SMAP L4 have been found to underestimate soil moisture within the Little Washita Watershed network[8,59] in Oklahoma, USA, which contains geophysical conditions (annual precipitation, land cover, terrain, etc.) similar to that of the HRB.

The differences in the temporal sampling period lead to potentially unfair comparisons between SMAP L3 and RF-SMAP. To examine this possibility, Figure 8 shows the fluctuations of the daily average of SMAP L3 retrieval frequency and IMERG precipitation for the pixel where the *in situ* sites locate. The time-domain distribution is such that SMAP L3 tends to collect relatively more retrievals during the wet season than in the dry season. That is to say, the gap-filling model is actively applied to fill gaps during the dry season

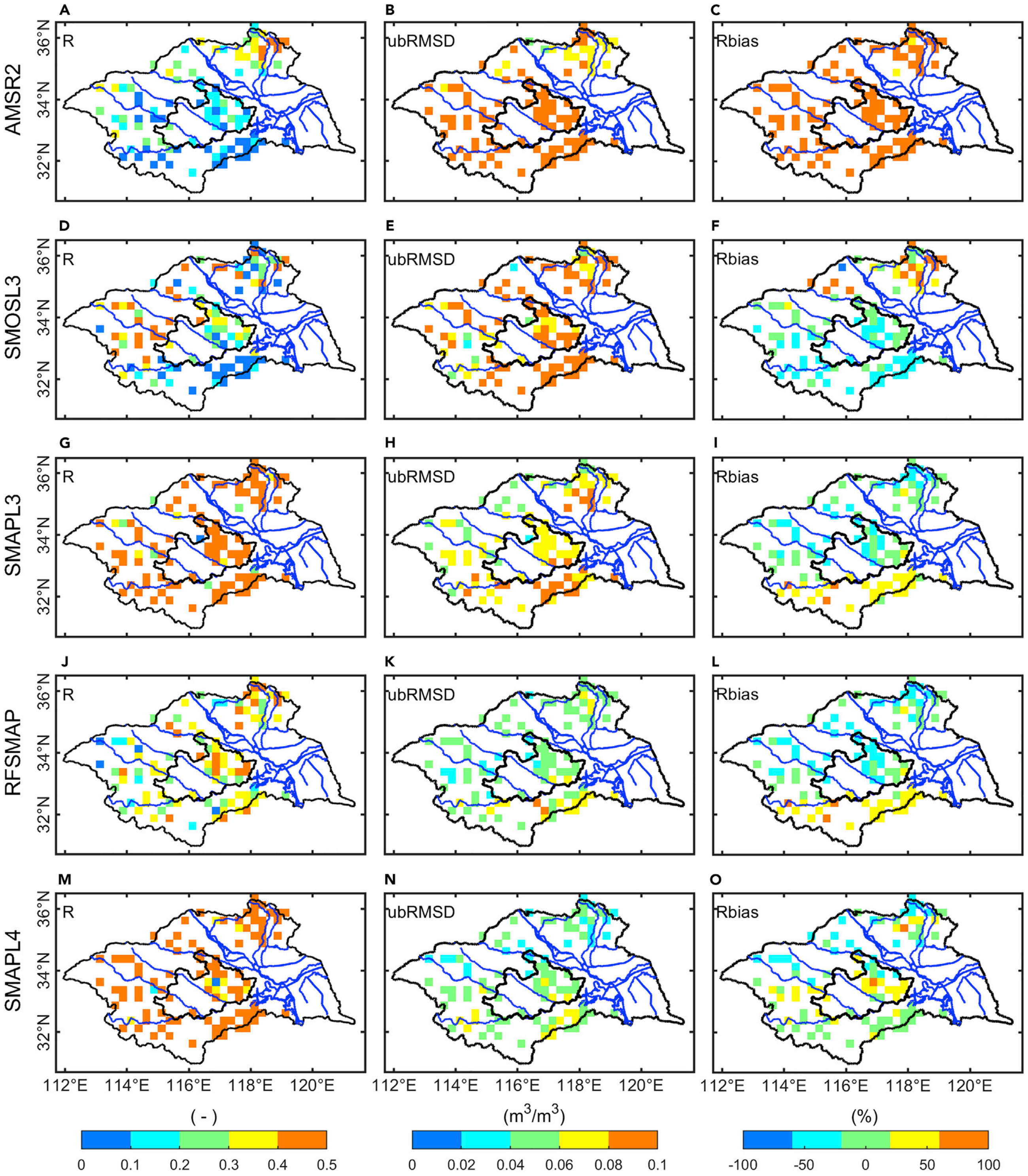


Figure 7. Statistical metrics based on in-situ observations for the RF-SMAP and four competitive datasets

(A–O) Ground observation-based temporal R, ubRMSD, and Rbias maps for AMSR2 (A–C), SMOS L3 (D–F), SMAP L3 (G–I), RF-SMAP (J–L), and SMAP L4 (M–O).

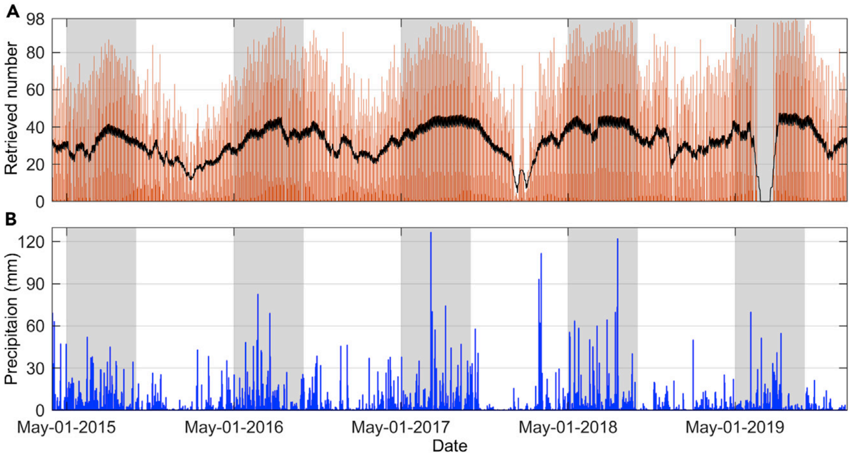


Figure 8. Time-series comparison of the retrieved quantities for SMAP_L3 descending orbit

(A and B) Time series for the daily number of SMAP L3 descending retrievals (A) and IMERG precipitation (B) during the study period in the site-located pixels (98). The black line indicates the temporal smoothing by using a 15-day moving-average window. Gray shading represents the wet season.

with relatively more complex surface conditions (e.g., heterogeneous variations of soil freeze/thaw state of snow cover). Similarly, a significant fraction of missing SMAP L3 observations is associated with intense rainfall. Therefore, the RF-SMAP product is tasked to retrieve soil moisture under relatively difficult conditions.

TC validation

Given the very sparse *in situ* observations in the central and eastern HRB, TC-based validation was used as a pixel-by-pixel alternative way to evaluate four passive-related soil moisture products (AMSR2, SMOS L3, SMAP L3, and RF-SMAP). Note that SMAP L4 was not applied in the TC approach because it likely contains obvious cross-correlated errors with both modeled data and the SMAP L3 product, which violates a core TC assumption. Therefore, the "ASCAT-AMSR2-CLDAS," "ASCAT-SMOS L3-CLDAS," "ASCAT-SMAP L3-CLDAS," and "ASCAT-RF-SMAP-CLDAS" triplets were involved separately in TC analysis during the daytime (AMSR2 ascending orbit at 1:30 PM, SMOS ascending orbit at 6:00 AM, and SMAP descending orbit at 6:00 AM) and nighttime (AMSR2 descending orbit at 1:30 AM, SMOS descending orbit at 6:00 PM, and SMAP ascending orbit at 6:00 PM).

In general, there are no obvious differences between daytime and nighttime TC-based results (Figure 9), and the spatial patterns of the TC-based metrics are similar to gauge-based ones shown earlier in Figure 7. SMAP L3 still provides the best performance, with most R-values generally above 0.6. The spatial pattern of R for RF-SMAP is similar to that for SMAP L3 except with degraded performance. Large areas of the eastern HRB demonstrate low R values for both SMAP L3 and RF-SMAP.

The other L-band retrieved product, SMOS L3, presents a spatial R distribution similar to that of RF-SMAP but with an area of missing results within the southeast corner of the HRB and relatively low R (i.e., nearly half of the SMOS R map is less than 0.4). Similarly, AMSR2 demonstrates consistently poor correlation particularly during nighttime (R ranges from 0 to 0.4).

The spatial distribution of the TC-based ubRMSD for daytime and nighttime retrievals is shown in Figure 10. For all products, little variations are seen between daytime and nighttime results. AMSR2 and SMOS L3 show higher ubRMSD over the whole HRB than SMAP-related products. In addition, the two products appear to have an opposing spatial relationship such that pixels with higher ubRMSD for AMSR2 correspond to lower ones for SMOS L3 and vice versa. As for SMAP L3 and RF-SMAP, both demonstrate widespread low ubRMSD. However, the ubRMSD for RF-SMAP is generally lower than that for RF-SMAP (most western pixels are <0.04 m^3/m^3) and SMAP L3, especially in the southern and eastern HRB. This emphasizes the benefit of integrating RF-predicted values with the original SMAP L3 time series. In addition, it is noteworthy that a common area of extremely high ubRMSD ($\geq$0.08 m^3/m^3) in all data sets is seen in the southeastern HRB.

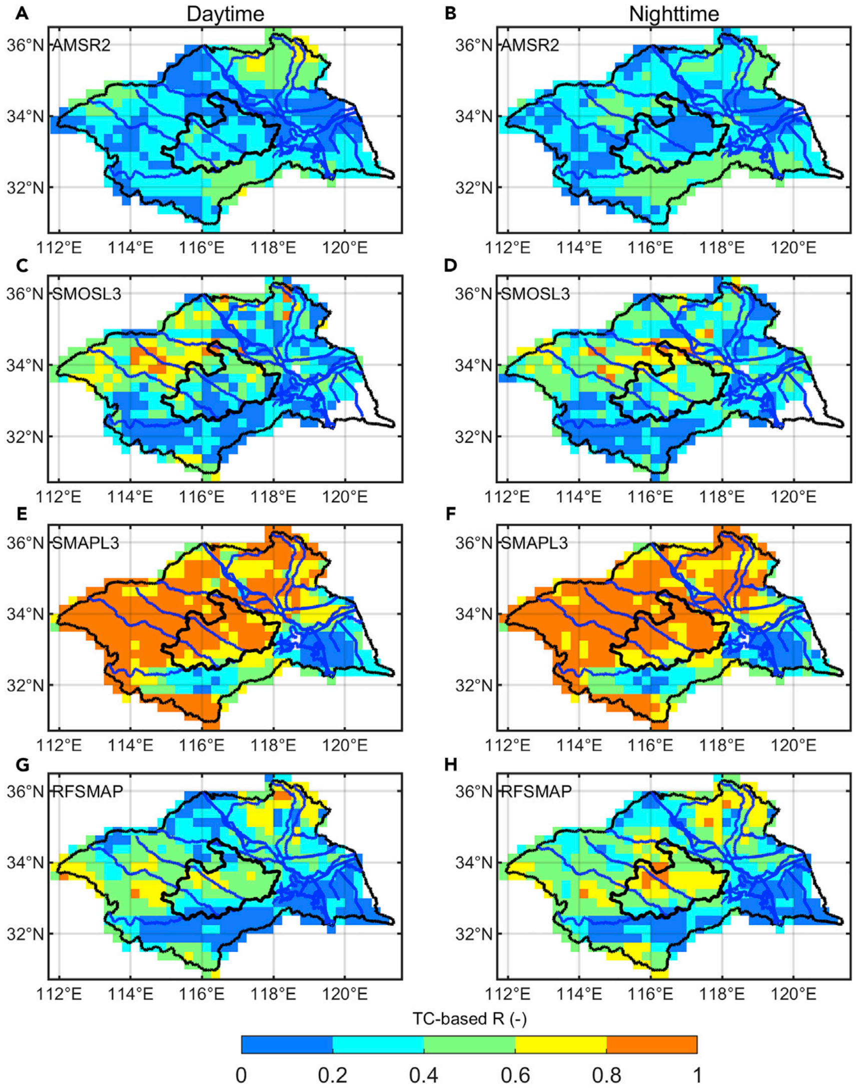


Figure 9. R-values map based on TC method for the RF-SMAP and four competitive datasets

(A–H) Temporal TC-based R maps for AMSR2 (A and B), SMOS L3 (C and D), SMAP L3 (E and F), and RF-SMAP (G and H) are calculated by single collocation triplet: "ASCAT-XXX-CLDAS," where "XXX" represents one of the passive remote sensing products (i.e., AMSR2, SMOS L3, or SMAP L3).

DISCUSSION

If supplied with good hyper-parameters, the proposed gap-filling model can be trained and run for 5-year records in a 270,000-km^2 area within a computational time of 15 minutes (platform: Intel Xeon W-2155 3.3 GHz/96 GB). As a result, the RF-SMAP product can be easily and quickly retrospectively updated to match new release versions of the SMAP L3 product. To ensure that this gap-filling pattern can be transmitted in an NRT system, these two key dynamic inputs (precipitation and LST) are collected from low-latency operational data products (i.e., global precipitation measurement products and GEOS-5 FP), and other ancillary statistic variables are consistent with the official data versions released by SMAP. The homogeneity of the area is a problem for SMAP, due to the Sun-synchronous orbit leaving gaps in data in these areas (Figures 2C and 2D). Therefore, this method can be readily extended to new geographic locations benefitted from the open-source inputs and programmatic hyper-parameters searching methods. Moreover, the test includes a total of 5-year data records (2015–2020) covering most of the time since the SMAP was launched, making the method robust in terms of temporal transferability.

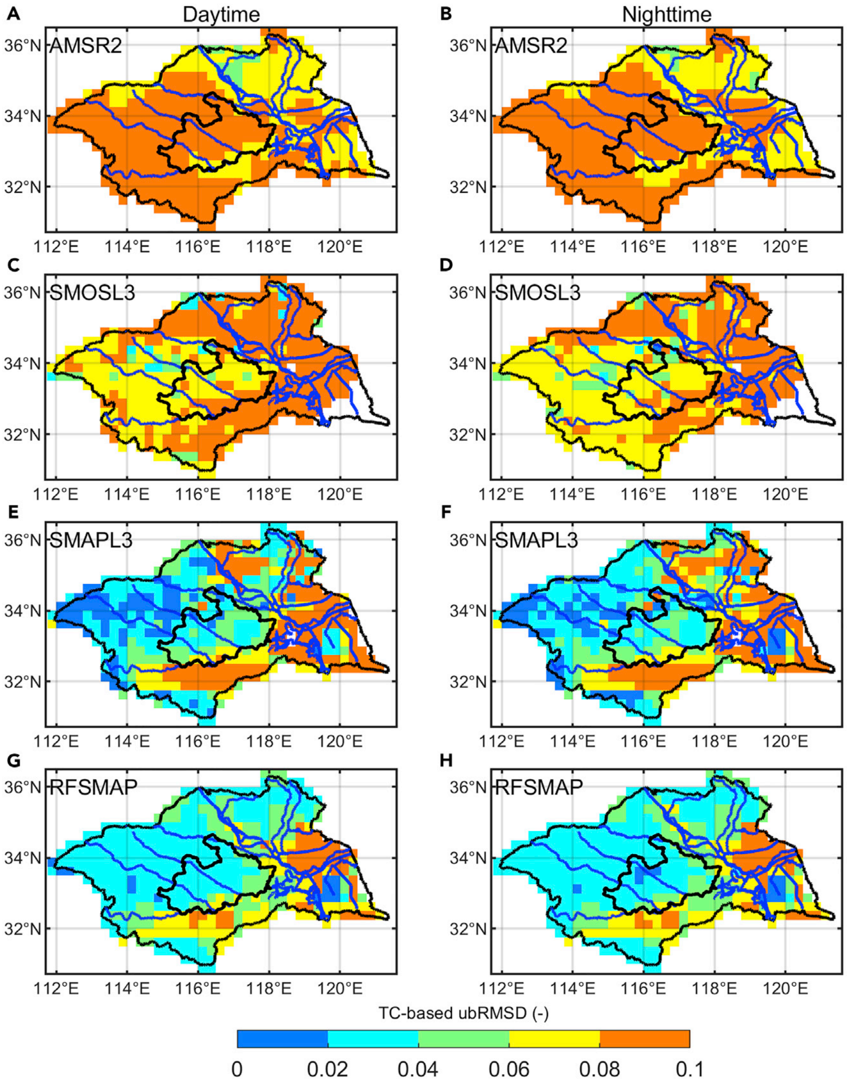


Figure 10. ubRMSD-values map based on TC method for the RF-SMAP and four competitive datasets
(A–H) Temporal TC-based ubRMSD maps for AMSR2 (A and B), SMOS L3 (C and D), SMAP L3 (E and F), and RF-SMAP (G and H) are calculated by single collocation triplet: "ASCAT-XXX-CLDAS", where "XXX" represents one of the passive remote sensing products (i.e., AMSR2, SMOS L3, or SMAP L3).

Although RF works reasonably well with the default hyper-parameters provided by software packages in most classification cases, the RF regression issues still require tuning the hyper-parameters to some extent for improving the model efficiency and accuracy.[60] Considering that the overall budget is very limited in most practical situations, the hyper-parameters optimization should be able to prioritize the evaluation of the objective function and have strong anytime performance, which indicates the ability to detect optimal or near-optimal configurations even with very limited budgets.[61] Given this, SHS is an extremely simple yet powerful multifidelity algorithm selection strategy especially to avoid significant budget consumption,[62] which can be easily called from scikit-learn to help users build a robust model.

In terms of the weight applied to specific inputs into the RF model (Table 4), relatively greater weights are assigned to static variables (e.g., the biggest contributor is latitude) versus time-varying variables. This phenomenon is closely connected with the climate characteristics of the HRB, where the warm temperate monsoon climate in the north and the subtropical climate in the south[63,64] typically lead to a larger amount of soil moisture spatial variability. Similarly, there is a gradual rising gradient for latitude from southeast to northwest. These two factors are intuitively reflected in latitude and DEM information.

Table 4. Features contribution score of RF internal training for ascending and descending orbit model

Features	Score-ascending (%)	Features	Score-descending (%)
Latitude	44.33	Latitude	42.99
DOY	12.61	DOY	13.21
DEM	11.71	DEM	11.56
LST	8.71	LST	10.63
Precipitation	6.72	Longitude	6.77
Longitude	6.48	Precipitation	4.30
Roughness	2.43	Roughness	3.47
Sand	2.30	Clay	2.22
Clay	2.25	Bulk density	1.80
Slope	1.28	Sand	1.74
Bulk density	1.19	Slope	1.30

In contrast, soil attributions (e.g., clay content, sand content, and bulk density) and surface descriptions (e.g., roughness, slope) have relatively little impact on prediction accuracy. This finding can be partly interpreted by the relatively flat terrain in the region and the relatively homogeneous soil type (lime concretion black soil and yellow fluvo-aquic soil) present in the HRB.[65,66] Despite the high contribution of static variables to model training, it can be seen from Figure 4 that the predicted data (RF-SMAP) still maintain adequate temporal variability in accordance with the time variation trend of the original SMAP L3.

Seasonal variations in the number of retrievals will be no doubt transmitted into the gap-filling model training, which means that the model has higher robustness in the case of a larger data size. For example, Figure 5 shows that a lower RF-SMAP quality can be found during the dry season than in the wet season. The consistency of such site validation can be compromised by the fact that there are very few valid reference values (involved in model training) during the winter season. However, the CDF of SMAP L3 is more consistent with the *in situ* data in the dry season than with those in the wet season, due to the relatively small temporal variation in soil moisture during the dry season, which allows for a good fit to the CDF. It was found that the number of SMAP L3 retrievals within the dry season increased slightly from year to year within the SMAP historical record. This implies that gap-filling will become progressively less difficult as the SMAP mission matures.

In situ validation and TC analysis can be mutually complementary. For example, because TC is blind to bias, Rbias values provided by *in situ* validation can complement TC-based results by providing a novel assessment of bias. In turn, TC-based metrics can provide precision-based evaluation metrics over continuous spatial domains—even in areas lacking *in situ* instrumentation. Note that although spatial patterns in temporal correlation (R) obtained from both *in situ* validation and TC analysis are roughly comparable (compare Figures 7 and 9), the TC-based correlations are higher. This is due to a few factors: (1) *In situ* validation is limited to available point-scale sites and, therefore, suffers from representativeness error when used to estimate grid-scale soil moisture,[67] whereas TC analysis generally compensates for this effect; (2) time and depth mismatches exist between fixed *in situ* measurements (8:00 AM, 0–10 cm) and satellite observation (C-band for 0–2 cm at 1:30 PM and L-band for 0–5 cm at 6:00 AM), which can complicate the interpretation of validation results; and (3) rigorous testing of TC assumptions is difficult, and the violation of these assumptions can bias TC-based results.

Under the recognition that both *in situ* validation and TC analysis have their insufficiency, some valuable evaluation information can still be obtained. Overall, all evaluated products show an acceptable correlation with true values in the area north of the Huai River mainstem, while the accuracy was very poor in the southeastern and southern HRB areas. However, in these areas, comparisons with *in situ* data suggest that RF-SMAP is less biased than other intercompared satellite products.

Intuitively, a key consideration is how much RF-SMAP outperforms other continuous products such as SMAP L4. While RF-SMAP retrievals are generally less precise than SMAP L3, and clearly inferior to

SMAP L4, they are also available at reduced temporal latency with an accuracy that is very close to the SMAP SM retrieval accuracy goal (i.e., ubRMSD = 0.04 m^3/m^3).

Conclusions

Benefitting from the NRT release of IMERG-E precipitation (~4-h latency) and GEOS-5 FP LST (~7-h latency), the proposed gap-filling model in this study can compensate the regional missing values of SMAP L3 within a few minutes once SMAP L3 releases, which is faster than the generation of SMAP L4 (a mean latency of ~2.5 days). This practice breaks through the limitations of application caused by the original data gaps and the high latency of the entire retrospective data products following the update of an operational algorithm.

Overall, RF-SMAP is able to follow the good performance of SMAP L3 very well. Its median ubRMSD is always around 0.05 m^3/m^3 compared to the *in situ* data, which is very close to the accuracy goal for the SMAP mission. Similarly, during the wet season, the median R for RF-SMAP (0.41) is higher than that for AMSR2 (0.17) and SMOS L3 (0.21) but lower than that for SMAP L3 (0.59) and SMAP L4 (0.57). This is because the proposed gap-filling model tends to predict soil moisture under conditions that are not conducive to satellite inversion (e.g., dense vegetation or intense rainfall, deterioration was observed for RF-SMAP during the dry season, which is due to a significant reduction in the sample size of SMAP L3 retrievals, required for RF-model training, during this period).

The TC results agree with *in situ*-based statistical metrics, suggesting that they can provide reliable evaluation results in the south and southeastern HRB where *in situ* observations are sparse. Poor R patterns can be observed for both RF-SMAP and SMAP L3 in these areas where the large manmade L-band RFI sources are reported and frequent precipitation events occur. Moreover, *in situ*-based R-values for the results are usually artificially degraded by upscaling/representativeness errors present in the ground.

Limitation of the study

Some limitations of this study are as follows: (1) The built model dominated by static inputs implicitly weakens the ability to predict extreme values because the model tends to increase accuracy with a certain value of certainty of predicted values; (2) the depth mismatch between *in situ* and satellite observations could bring potential uncertainty toward *in situ* validation results; and (3) despite the high accessibility of the data required by the model, the method has only been tested in relatively flat terrain, predominantly agricultural study area, and further research is needed to reconstruct other more complex areas.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Hardware and computing environment used

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.105853.

ACKNOWLEDGMENTS

This study is supported by the National Key Research and Development Program (no. 2018YFA0605400) and the National Natural Science Foundation of China (no. 41830752, 41961134003, and 42101397). The United States Department of Agriculture (USDA) is an open-opportunity employer. The findings, representations, and conclusions in this publication are those of the authors and should not be construed to represent any official USDA or US Government determination or policy.

AUTHOR CONTRIBUTIONS

H.L. and X.W. conceived and designed the research framework. J.S., J.Z., and X.W. collected and preprocessed the data. X.W., G.C., and W.T.C. optimized the model. X.W., H.L., and W.T.C. wrote the manuscript. G.C., Y.Z., Q.G., and H.L. supervised the entire study.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 19, 2022
Revised: November 9, 2022
Accepted: December 19, 2022
Published: January 20, 2023

REFERENCES

1. Daganzo-Eusebio, E., Oliva, R., Kerr, Y.H., Nieto, S., Richaume, P., and Mecklenburg, S.M. (2013). SMOS radiometer in the 1400–1427-MHz passive band: impact of the RFI environment and approach to its mitigation and cancellation. IEEE Trans. Geosci. Rem. Sens. *51*, 4999–5007. https://doi.org/10.1109/TGRS.2013.2259179.

2. Dharssi, I., Bovis, K.J., Macpherson, B., and Jones, C.P. (2011). Operational assimilation of ASCAT surface soil wetness at the. Hydrol. Earth Syst. Sci. *15*, 2729–2746. https://doi.org/10.5194/hess-15-2729-2011.

3. Holzman, M.E., Carmona, F., Rivas, R., and Niclòs, R. (2018). Early assessment of crop yield from remotely sensed water stress and solar radiation data. ISPRS J. Photogrammetry Remote Sens. *145*, 297–308. https://doi.org/10.1016/j.isprsjprs.2018.03.014.

4. Mladenova, I.E., Bolten, J.D., Crow, W.T., Anderson, M.C., Hain, C.R., Johnson, D.M., and Mueller, R. (2017). Intercomparison of soil moisture, evaporative stress, and vegetation indices for estimating corn and soybean yields over the U.S. IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. *10*, 1328–1343. https://doi.org/10.1109/JSTARS.2016.2639338.

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

6. Engman, E.T., and Chauhan, N. (1995). Status of microwave soil moisture measurements with remote sensing. Rem. Sens. Environ. *51*, 189–198. https://doi.org/10.1016/0034-4257(94)00074-W.

7. Escobar, V.M., Srinivasan, M., and Arias, S.D. (2016). Improving NASA's Earth observation systems and data programs through the engagement of mission early adopters. In Earth Science Satellite Applications: Current and Future Prospects, F. Hossain, ed. (Springer International Publishing), pp. 223–267. https://doi.org/10.1007/978-3-319-33438-7_9.

8. Chen, Q., Zeng, J., Cui, C., Li, Z., Chen, K.S., Bai, X., and Xu, J. (2018). Soil moisture retrieval from SMAP: a validation and error analysis study using ground-based observations over the little Washita watershed. IEEE Trans. Geosci. Rem. Sens. *56*, 1394–1408. https://doi.org/10.1109/TGRS.2017.2762462.

9. Stillman, S., and Zeng, X. (2018). Evaluation of SMAP soil moisture relative to five other satellite products using the climate reference network measurements over USA. IEEE Trans. Geosci. Rem. Sens. *56*, 6296–6305. https://doi.org/10.1109/tgrs.2018.2835316.

10. Tavakol, A., Rahmani, V., Quiring, S.M., and Kumar, S.V. (2019). Evaluation analysis of NASA SMAP L3 and L4 and SPoRT-LIS soil moisture data in the United States. Rem. Sens. Environ. *229*, 234–246. https://doi.org/10.1016/j.rse.2019.05.006.

11. Wigneron, J.P., Jackson, T.J., O'Neill, P., De Lannoy, G., de Rosnay, P., Walker, J.P., Ferrazzoli, P., Mironov, V., Bircher, S., Grant, J.P., et al. (2017). Modelling the passive microwave signature from land surfaces: a review of recent results and application to the L-band SMOS & SMAP soil moisture retrieval algorithms. Rem. Sens. Environ. *192*, 238–262. https://doi.org/10.1016/j.rse.2017.01.024.

12. Wu, H., Kimball, J.S., Zhou, N., Alfieri, L., Luo, L., Du, J., and Huang, Z. (2019). Evaluation of real-time global flood modeling with satellite surface inundation observations from SMAP. Rem. Sens. Environ. *233*, 111360. https://doi.org/10.1016/j.rse.2019.111360.

13. Lawston, P. M., Santanello, J. A., and Kumar, S. V. (2017). Irrigation signals detected from SMAP soil moisture retrievals. *Geophysical Research Letters*, 44, 11860–11867. https://doi.org/10.1002/2017GL075733

14. Mladenova, I.E., Bolten, J.D., Crow, W., Sazib, N., and Reynolds, C. (2020). Agricultural drought monitoring via the assimilation of SMAP soil moisture retrievals into a global soil water balance model. Front. Big Data *3*, 10. https://doi.org/10.3389/fdata.2020.00010.

15. Reichle, R.H., De Lannoy, G.J.M., Liu, Q., Ardizzone, J.V., Colliander, A., Conaty, A., Crow, W., Jackson, T.J., Jones, L.A., Kimball, J.S., et al. (2017). Assessment of the SMAP level-4 surface and root-zone soil moisture product using in situ measurements. J. Hydrometeorol. *18*, 2621–2645. https://doi.org/10.1175/jhm-d-17-0063.1.

16. Reichle, R.H., De Lannoy, G.J.M., Liu, Q., Koster, R.D., Kimball, J.S., Crow, W.T., Ardizzone, J.V., Chakraborty, P., Collins, D.W., Conaty, A.L., et al. (2017). Global assessment of the SMAP level-4 surface and root-zone soil moisture product using assimilation diagnostics. J. Hydrometeorol. *18*, 3217–3237. https://doi.org/10.1175/jhm-d-17-0130.1.

17. Liu, Y., Yang, Y., and Jing, W. (2020). Potential applicability of SMAP in ECV soil moisture gap-filling: a case study in europe. IEEE Access *8*, 133114–133127. https://doi.org/10.1109/ACCESS.2020.3009977.

18. Kim, H., and Lakshmi, V. (2018). Use of cyclone global navigation satellite system (CyGNSS) observations for estimation of soil moisture. Geophys. Res. Lett. *45*, 8272–8282. https://doi.org/10.1029/2018GL078923.

19. Gruber, A., Dorigo, W.A., Crow, W., and Wagner, W. (2017). Triple collocation-based merging of satellite soil moisture retrievals. IEEE Trans. Geosci. Rem. Sens. *55*, 6780–6792. https://doi.org/10.1109/TGRS.2017.2734070.

20. Zakeri, F., and Mariethoz, G. (2021). A review of geostatistical simulation models applied to satellite remote sensing: methods and applications. Rem. Sens. Environ. *259*, 112381. https://doi.org/10.1016/j.rse.2021.112381.

21. Llamas, R.M., Guevara, M., Rorabaugh, D., Taufer, M., and Vargas, R. (2020). Spatial gap-filling of ESA CCI satellite-derived soil moisture based on geostatistical techniques and multiple regression. Rem. Sens. *12*, 665.

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

23. Sun, H., and Xu, Q. (2021). Evaluating machine learning and geostatistical methods

for spatial gap-filling of monthly ESA CCI soil moisture in China. Rem. Sens. *13*, 2848.

24. Tong, C., Wang, H., Magagi, R., Goïta, K., and Wang, K. (2021). Spatial gap-filling of SMAP soil moisture pixels over Tibetan plateau via machine learning versus geostatistics. IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. *14*, 9899–9912. https://doi.org/10.1109/JSTARS.2021.3112623.

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

26. Long, D., Bai, L., Yan, L., Zhang, C., Yang, W., Lei, H., et al. (2019). Generation of spatially complete and daily continuous surface soil moisture of high spatial resolution. Remote Sens. Environ. *233*, 111364.

27. Baker, B., Gupta, O., Raskar, R., and Naik, N. (2017). Accelerating neural architecture search using performance prediction. Preprint at arXiv. https://doi.org/10.48550/arXiv.1705.10823.

28. Chao, G. (2010). Analysis and Research on Causes of Climat-Hydrological Change in the Huaihe River Basin. Doctor of Philosophy (Nanjing Institute of Geography and Limnolog of Chinese Academy of Sciences).

29. Wang, X., Lü, H., Crow, W.T., Zhu, Y., Wang, Q., Su, J., Zheng, J., and Gou, Q. (2021). Assessment of SMOS and SMAP soil moisture products against new estimates combining physical model, a statistical model, and in-situ observations: a case study over the Huai River Basin, China. J. Hydrol. *598*, 126468. https://doi.org/10.1016/j.jhydrol.2021.126468.

30. Hong, G., Yongsheng, J., Deyi, C., Jinbiao, X., Jianzhong, M., Liu, G., Mingkai, Q., Xushui, C., Yongfa, F., Chaohui, Y., et al. (2020). Huai River water resources bulletin (2019). In The Huaihe River Commission of the Ministry of Water Resources, P.R.C., ed.

31. Huffman, G.J., Stocker, E.F., Bolvin, D.T., Nelkin, E.J., and Tan, J. (2019). In GPM IMERG Early Precipitation L3 Half Hourly 0.1 degree x 0.1 degree V06, G.E.S.D.a.I.S.C.G. DISC, ed.

32. Yu, C., Hu, D., Di, Y., and Wang, Y. (2021). Performance evaluation of IMERG precipitation products during typhoon Lekima (2019). J. Hydrol. *597*, 126307. https://doi.org/10.1016/j.jhydrol.2021.126307.

33. Tang, G., Zeng, Z., Ma, M., Liu, R., Wen, Y., and Hong, Y. (2017). Can near-real-time satellite precipitation products capture rainstorms and guide flood warning for the 2016 summer in South China? Geosci. Rem. Sens. Lett. IEEE *14*, 1208–1212.

34. Peng, J., Mohammed, P., Chaubell, J., Chan, S., Kim, S., Das, N., Dunbar, S., Bindlish, R., and Xu, X. (2019). Soil Moisture Active Passive (SMAP) L1-L3 Ancillary GEOS Data, Version 1, First edition (National Snow and Ice Data Center).

35. Zhang, s., Wang, G., Yu, p., Zou, w., Jin, f., Zhao, h., Cui, x., Zhang, s., Gu, l., Liu, m., et al. (2015). Specifications for Soil Moisture Monitoring (In Chinese) (China Water&Power Press).

36. Wagner, W., Hahn, S., Kidd, R., Melzer, T., Bartalis, Z., Hasenauer, S., Figa-Saldaña, J., De Rosnay, P., Jann, A., Schneider, S., et al. (2013). The ASCAT soil moisture product: a review of its specifications, validation results, and emerging applications. metz. *22*, 5–33.

37. Sadikni, R., and Jahnke-Bornemann, A. (2021). ASCAT Global Maps of daily running 5-day mean surface soil moisture. In 2020_fv0.01, I.C.D.C. (ICDC), ed..

38. Kerr, Y.H., Al-Yaari, A., Rodriguez-Fernandez, N., Parrens, M., Molero, B., Leroux, D., Bircher, S., Mahmoodi, A., Mialon, A., Richaume, P., et al. (2016). Overview of SMOS performance in terms of global soil moisture monitoring after six years in operation. Rem. Sens. Environ. *180*, 40–63. https://doi.org/10.1016/j.rse.2016.02.042.

39. Al Bitar, A., Mialon, A., Kerr, Y.H., Cabot, F., Richaume, P., Jacquette, E., Quesney, A., Mahmoodi, A., Tarot, S., Parrens, M., et al. (2017). The global SMOS Level 3 daily soil moisture and brightness temperature maps. Earth Syst. Sci. Data *9*, 293–315. https://doi.org/10.5194/essd-9-293-2017.

40. Kim, S., Liu, Y., Johnson, F.M., Parinussa, R.M., and Sharma, A. (2015). A global comparison of alternate AMSR2 soil moisture products: why do they differ? Rem. Sens. Environ. *161*, 43–62. https://doi.org/10.1016/j.rse.2015.02.002.

41. Wu, Q., Liu, H., Wang, L., and Deng, C. (2016). Evaluation of AMSR2 soil moisture products over the contiguous United States using in situ data from the International Soil Moisture Network. Int. J. Appl. Earth Obs. Geoinf. *45*, 187–199. https://doi.org/10.1016/j.jag.2015.10.011.

42. Jeu, R.d., and Owe, M. (2014). In AMSR2/GCOM-W1 surface soil moisture (LPRM) L3 1 day 25 km x 25 km V001, G.E.S.D.a.I.S.C.G. DISC, ed.

43. Shi, C., Jiang, L., Zhang, T., Xu, B., and Han, S. (2014). Status and Plans of CMA Land Data Assimilation System (CLDAS) Project, p. 5671.

44. Jiang, Z., Shi, C., Han, S., Zhang, T., and Zhu, Z. (2015). CLDAS Atmospheric Forcing Field Products V2.0.

45. Reichle, R., Lannoy, G.D., Koster, R.D., Crow, W.T., Kimball, J.S., and Liu, Q. (2020). SMAP L4 Global 3-hourly 9 km EASE-Grid Surface and Root Zone Soil Moisture Geophysical Data, Version 5. [Indicate subset used] (National Snow and Ice Data Center).

46. Reichle, R.H., Ardizzone, J.V., Kim, G.-K., Lucchesi, R.A., Smith, E.B., and Weiss, B.H. (2018). Soil Moisture Active Passive (SMAP) Mission Level 4 Surface and Root Zone Soil Moisture (L4_SM) Product Specification Document.

47. Reichle, R., Koster, R., De Lannoy, G., Crow, W., and Kimball, J. (2014). Level 4 Surface and Root Zone Soil Moisture (L4_SM) Data Product.

48. Al-Yaari, A., Wigneron, J.P., Ducharne, A., Kerr, Y., de Rosnay, P., de Jeu, R., Govind, A., Al Bitar, A., Albergel, C., Muñoz-Sabater, J., et al. (2014). Global-scale evaluation of two satellite-based passive microwave soil moisture datasets (SMOS and AMSR-E) with respect to Land Data Assimilation System estimates. Rem. Sens. Environ. *149*, 181–195. https://doi.org/10.1016/j.rse.2014.04.006.

49. Breiman, L. (2001). Random forests. Mach. Learn. *45*, 5–32. https://doi.org/10.1023/A:1010933404324.

50. Oshiro, T.M., Perez, P.S., and Baranauskas, J.A. (2012). How Many Trees in a Random Forest? Held (Springer Berlin Heidelberg)), pp. 154–168.

51. Hutter, F., Kotthoff, L., and Vanschoren, J. (2019). Automated Machine Learning: Methods, Systems, Challenges (Springer Nature).

52. Yang, L., and Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: theory and practice. Neurocomputing *415*, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061.

53. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

54. Stoffelen, A. (1998). Toward the true near-surface wind speed: error modeling and calibration using triple collocation. J. Geophys. Res. *103*, 7755–7766. https://doi.org/10.1029/97JC03180.

55. Gruber, A., Su, C.H., Zwieback, S., Crow, W., Dorigo, W., and Wagner, W. (2016). Recent advances in (soil moisture) triple collocation analysis. Int. J. Appl. Earth Obs. Geoinf. *45*, 200–211. https://doi.org/10.1016/j.jag.2015.09.002.

56. McColl, K.A., Vogelzang, J., Konings, A.G., Entekhabi, D., Piles, M., and Stoffelen, A. (2014). Extended triple collocation: estimating errors and correlation coefficients with respect to an unknown target. Geophys. Res. Lett. *41*, 6229–6236. https://doi.org/10.1002/2014GL061322.

57. Soldo, Y., Le Vine, D.M., de Matthaeis, P., and Richaume, P. (2017). L-band RFI detected by SMOS and aquarius. IEEE Trans. Geosci. Rem. Sens. *55*, 4220–4235. https://doi.org/10.1109/TGRS.2017.2690406.

58. Konings, A.G., Piles, M., Das, N., and Entekhabi, D. (2017). L-band vegetation optical depth and effective scattering albedo estimation from SMAP. Rem. Sens. Environ. *198*, 460–470. https://doi.org/10.1016/j.rse.2017.06.037.

59. Cui, C., Xu, J., Zeng, J., Chen, K.-S., Bai, X., Lu, H., Chen, Q., and Zhao, T. (2017). Soil moisture mapping from satellites: an intercomparison of SMAP, SMOS, FY3B, AMSR2, and ESA CCI over two dense network regions at different spatial scales. Rem. Sens. *10*, 33.

60. Probst, P., Wright, M.N., and Boulesteix, A. (2019). Hyperparameters and tuning strategies for random forest. WIREs Data Mining Knowl. Discov. *9*, e1301. https://doi.org/10.1002/widm.1301.

61. Falkner, S., Klein, A., and Hutter, F. (2018). BOHB: robust and efficient hyperparameter optimization at scale. In Proceedings of the 35th International Conference on Machine Learning, D. Jennifer and K. Andreas, eds. (PMLR).

62. Feurer, M., and Hutter, F. (2019). Hyperparameter optimization. In Automated machine learning (Springer), pp. 3–33.

63. He, Y., Ye, J., and Yang, X. (2015). Analysis of the spatio-temporal patterns of dry and wet conditions in the Huai River Basin using the standardized precipitation index. Atmos. Res. *166*, 120–128. https://doi.org/10.1016/j.atmosres.2015.06.022.

64. Li, M., Chu, R., Shen, S., and Islam, A.R.M.T. (2018). Dynamic analysis of pan evaporation variations in the Huai River Basin, a climate transition zone in eastern China. Sci. Total Environ. *625*, 496–509. https://doi.org/10.1016/j.scitotenv.2017.12.317.

65. Zhai, L., Xu, P., Zhang, Z., Li, S., Xie, R., Zhai, L., and Wei, B. (2017). Effects of deep vertical rotary tillage on dry matter accumulation and grain yield of summer maize in the Huang-Huai-Hai Plain of China. Soil Tillage Res. *170*, 167–174.

66. Mingcheng, D., Zhenlong, W., Cuiling, J., Faxin, W., and Chao, Z. (2018). Simulation of runoff and sediment production regularity of different rainfall intensity and changeable slope gradients in the yellow fluvo-aquic soil oh the huaibei plain (in Chinese). J. Soil Water Conserv. *32*, 6.

67. Chen, F., Crow, W.T., Cosh, M.H., Colliander, A., Asanuma, J., Berg, A., Bosch, D.D., Caldwell, T.G., Collins, C.H., Jensen, K.H., et al. (2019). Uncertainty of reference pixel soil moisture averages sampled at SMAP core validation sites. J. Hydrometeorol. *20*, 1553–1569. https://doi.org/10.1175/jhm-d-19-0049.1.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
Python version 3.9	Python Software Foundation	https://www.python.org/
MATLAB (Figures plotting)	MathWorks	https://www.mathworks.com/
ArcGIS 10.8 (study area)	ESRI	https://support.esri.com/en/products/desktop/arcgis-desktop/arcmap/10-8
Scikit-learn (RF Model)	Google	https://scikit-learn.org/
Snellius	Dutch National Supercomputer	https://www.surf.nl/en/dutch-national-supercomputer-snellius
Deposited data		
SMAP_L3 soil moisture	NASA	https://nsidc.org/data/spl3smp/versions/8#anchor-1
SMAP_L3 Ancillary Data (GEOS-5 FP LST is included)	NASA/GMAO	https://nsidc.org/data/SMAP_L1_L3_ANC_GEOS/versions/1
SMOS	ESA	https://bec.icm.csic.es/global-land-datasets/
ASCAT	NOAA	https://www.cen.uni-hamburg.de/en/icdc/data/land/ascat-soilmoisture.html
AMSR2	GES DISC	https://disc.gsfc.nasa.gov/datasets/LPRM_AMSR2_DS_A_SOILM3_001/summary
CLDAS	CMA	http://data.cma.cn/mdrd/?r=data/detail&dataCode=NAFP_CLDAS2.0_NRT

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Xiaoyi Wang (wangxiaoyi.nk@gmail.com).

Materials availability

This study did not generate new unique physical materials.

Data and code availability

- The download links of all involved datasets are listed in the key resources table.
- The key codes have been deposited at GitHub (https://github.com/xiaoyi-wong/RF-SMAP) and are publicly available as of the date of publication.
- Any detailed information about this paper is available from the lead contact upon request.

METHOD DETAILS

In the data pre-processing stage, quality control was not applied to SMAP_L3 in this paper, this is given the dramatic reduction in the number of training targets after quality control and the small difference in performance before and after data quality (see Figure S1). Similarly, all the competing datasets omitted quality control procedure for the sake of fairness. Besides, a simple correlation analysis among variables involved in gap-filling modeling has been shown in Figure S2, which could be useful for the influence of each variable in prediction values and the accuracy of gap-filling model.

The detailed description of the random forest run setup and the derivation of the TC method has been listed as follow.

For the Random Forest regressor, all the datasets were acquired from March 31, 2015, to December 31, 2019, and resampled at a geographic grid with a 0.25-degree resolution by the nearest interpolation method.

In order to match the requirement of the RF model, the 3D (date-latitude-longitude) inputs and target data containing were forced to be flattened as a 2D array (date-variables). The dataset would be divided into two parts according to whether the target is marked as missing or not, respectively the successful retrievals set and the missing set waiting to be filled. The former could be fed into the RF model to forcefully find the relationship between the inputs and the target, the trained model could be driven by the latter data corresponding to the missing set. After that, the output would be reshaped into a 3D array assigned to the locations where SMAP fails to measure.

Hyperparameter tuning consists of having successive random search (HRS) and successive grid search (HGS), which is much less time-consuming than traditional methods (e.g., learning curve and grid search). In addition to the discrete parameters listed in Table 3, the search steps of n_estimators and max_depth set in HRS and HGS are 200 and 10, respectively. At the same time, enabling parallel computing (n_job = −1, a key parameter in scikit-learn) can significantly improve the training efficiency of the model. Note that the initial best n_estimators for Ascending model is 1600, and its output model is very large (~50 GB), so we set it to 1200 instead (the resulting training accuracy is barely affected), and then the prediction intervals of the gap-filling model can be seen in Figure S3.

For Triple Collocation (TC), its detailed derivation is summarized as follows:

$$\begin{aligned} X &= \alpha_X + \beta_X \cdot T + \varepsilon_X \\ Y &= \alpha_Y + \beta_Y \cdot T + \varepsilon_Y \\ Z &= \alpha_Z + \beta_Z \cdot T + \varepsilon_Z \end{aligned} \qquad \text{(Equation A.1)}$$

Where X, Y, and Z represent the observation from the non-homologous platform, T is the unknown true soil moisture, α and β denote the systematic additive and multiplicative biases of the dataset (X or Y or Z). Their covariance between random every two triplets can be calculated as:

$$\begin{aligned} Q_{XY} &= Cov(X,Y) \\ &= \beta_X\beta_Y\sigma_T^2 + \beta_X Cov(\varepsilon_Y,T) + \beta_Y Cov(\varepsilon_X,T) + \alpha_X E(\varepsilon_Y) \\ &\quad + \alpha_Y E(\varepsilon_X) + Cov(\varepsilon_X,\varepsilon_Y) \\ Q_{XZ} &= Cov(X,Z) \\ &= \beta_X\beta_Z\sigma_T^2 + \beta_X Cov(\varepsilon_Z,T) + \beta_Z Cov(\varepsilon_X,T) + \alpha_X E(\varepsilon_Z) \\ &\quad + \alpha_Z E(\varepsilon_X) + Cov(\varepsilon_X,\varepsilon_Z) \\ Q_{YZ} &= Cov(Y,Z) \\ &= \beta_Y\beta_Z\sigma_T^2 + \beta_Y Cov(\varepsilon_Z,T) + \beta_Z Cov(\varepsilon_Y,T) + \alpha_Y E(\varepsilon_Z) \\ &\quad + \alpha_Z E(\varepsilon_Y) + Cov(\varepsilon_Y,\varepsilon_Z) \end{aligned} \qquad \text{(Equation A.2)}$$

Where Q and E are covariance and expectation, $\sigma^2(\cdot)$ denotes the variance. Under the assumptions of the TC method: a). the random errors of each triplet have zero expectation, i.e., $E(\varepsilon) = 0$; b). the random errors of each triplet are independent of the true value, i.e., $Cov(\varepsilon_i, T), i \in (X, Y, Z)$; c). the random errors of each triplet are independent of each other, i.e., $Cov(i,j) = 0, i \neq j$ Equation A.2 can be written as:

$$Q_{i,j} = \begin{cases} \beta_i\beta_j\sigma_T^2(T) + \sigma^2(\varepsilon), i = j \\ \beta_i\beta_j\sigma_T^2(T), i \neq j \\ i,j \in (X,Y,Z) \end{cases} \qquad \text{(Equation A.3)}$$

By solving for the variance from Equation A.3, ubRMSD can be calculated as below:

$$\begin{aligned} ubRMSD_X &= \sqrt{\sigma^2(\varepsilon_X)} = \sqrt{Q_{XX} - \frac{Q_{XY}Q_{XZ}}{Q_{YZ}}} \\ ubRMSD_Y &= \sqrt{\sigma^2(\varepsilon_Y)} = \sqrt{Q_{YY} - \frac{Q_{XY}Q_{YZ}}{Q_{XZ}}} \\ ubRMSD_Z &= \sqrt{\sigma^2(\varepsilon_Z)} = \sqrt{Q_{ZZ} - \frac{Q_{ZX}Q_{ZY}}{Q_{XY}}} \end{aligned} \qquad \text{(Equation A.4)}$$

Furthermore, for acquiring the correlation estimate, $\beta_i\sigma_t$ can be regarded as θ_i the covariance is converted to:

$$Q_{i,j} = \begin{cases} \theta_i^2 + \sigma_{\varepsilon_j}^2, i = j \\ \theta_i \theta_j, i \neq j \\ i, j \in (X, Y, Z) \end{cases} \tag{Equation A.5}$$

The correlation coefficient ρ is obtained by deducing the mathematical relationship between the ρ and the internal variables of TC:

$$R_{i,T} = \frac{Cov(i,T)}{\sqrt{\sigma^2(i) \bullet \sigma^2(T)}} = \frac{\beta_i \sigma_T^2}{\sqrt{Q_{ii} \bullet \sigma^2(T)}} = \frac{\theta}{\sqrt{Q_{ii}}} \tag{Equation A.6}$$

Therefore, the correlation coefficients for each dataset against "T" are as follows

$$\begin{aligned} R_X &= \sqrt{\frac{Q_{XY} Q_{XZ}}{Q_{XX} Q_{YZ}}} \\ R_Y &= \sqrt{\frac{Q_{YX} Q_{YZ}}{Q_{YY} Q_{XY}}} \\ R_Z &= \sqrt{\frac{Q_{ZX} Q_{ZY}}{Q_{ZZ} Q_{XY}}} \end{aligned} \tag{Equation A.7}$$

Hardware and computing environment used

In this research, the study area is made by ArcGIS 10.8, the data pre-processing is operated by Python and MATLAB. The Random Forest model is built by Python 3.9 (the machine learning module is driven by scikit-learn 1.1.2). All the analysis (statistical metrics and TC analysis) and plots are performed in MATLAB 2021a. Apart from that, the High-Performance Computer (HPC) platform – Snellius of the Netherlands provides great help in accelerating program handling.