remote sensing

Article

Integration of Satellite-Derived and Ground-Based Soil Moisture Observations for a Precipitation Product over the Upper Heihe River Basin, China

Ying Zhang, Jinliang Hou * and Chunlin Huang

Key Laboratory of Remote Sensing of Gansu Province, Heihe Remote Sensing Experimental Research Station, Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou 730000, China
* Correspondence: jlhours@lzb.ac.cn

Abstract: Precipitation monitoring is important for earth system modeling and environmental management. Low spatial representativeness limits gauge measurements of rainfall and low spatial resolution limits satellite-derived rainfall. SM2RAIN-based products, which exploit the inversion of the water balance equation to derive rainfall from soil moisture (SM) observations, can be an alternative. However, the quality of SM data limits the accuracy of rainfall. The goal of this work was to improve the accuracy of rainfall estimation through merging multiple soil moisture (SM) datasets. This study proposed an integration framework, which consists of multiple machine learning methods, to use satellite and ground-based soil moisture observations to derive a precipitation product. First, three machine learning (ML) methods (random forest (RF), long short-term memory (LSTM), and convolutional neural network (CNN)) were used, respectively to generate three SM datasets (RF-SM, LSTM-SM, and CNN-SM) by merging satellite (SMOS, SMAP, and ASCAT) and ground-based SM observations. Then, these SM datasets were merged using the Bayesian model averaging method and validated by wireless sensor network (WSN) observations. Finally, the merged SM data were used to produce a rainfall dataset (SM2R) using SM2RAIN. The SM2R dataset was validated using automatic meteorological station (AMS) rainfall observations recorded throughout the Upper Heihe River Basin (China) during 2014–2015 and compared with other rainfall datasets. Our results revealed that the quality of the SM2R data outperforms that of GPM-SM2RAIN, Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS), ERA5-Land (ERA5) and multi-source weighted-ensemble Precipitation (MSWEP). Triple-collocation analysis revealed that SM2R outperformed China Meteorological Data and the China Meteorological Forcing Dataset. Ultimately, the SM2R rainfall product was considered successful with acceptably low spatiotemporal errors (RMSE = 3.5 mm, R = 0.59, and bias = −1.6 mm).

Keywords: soil moisture; rainfall; SM2RAIN; SMAP; SMOS; ASCAT

check for updates

Citation: Zhang, Y.; Hou, J.; Huang, C. Integration of Satellite-Derived and Ground-Based Soil Moisture Observations for a Precipitation Product over the Upper Heihe River Basin, China. *Remote Sens.* **2022**, *14*, 5355. https://doi.org/10.3390/rs14215355

Academic Editors: Yuei-An Liou and Arun Mondal

Received: 15 September 2022
Accepted: 22 October 2022
Published: 26 October 2022

1. Introduction

Monitoring of precipitation is essential for earth system modeling and environmental management, such as flood and landslide forecasting and water resources management. Precipitation can be derived from gauge stations, ground-based weather radars, and satellite radars and radiometers [1]. With the steady development of remote sensing, rainfall retrievals from satellite have proven successful and become increasingly important in various applications globally [2], e.g., meteorology, hydrology, agricultural applications, drought monitoring, flood forecasting, and water resources management [3–5].

Ground-based measurements of rainfall, which are generally considered the most accurate. However, the spatial and temporal coverages they provide are inadequate for many applications [6]. Moreover, the spatial distribution of rain gauges is often insufficiently dense, and the instruments are often not well maintained by local governments or organizations in many remote rural areas around the word, especially in developing countries. Generally, rain gauge networks are most inadequate in areas where they are needed most. Low spatial representativeness limits applications of gauge measurements. Fortunately, rainfall information for remote rural areas can be derived from remote sensing data [7–10]. Satellite rainfall is usually derived using microwave, visible, or infrared sensors onboard low-Earth-orbiting and geostationary satellites. Visible and infrared sensors are generally deployed on geostationary satellites to retrieve rainfall signals with high spatial and low temporal resolutions. Passive microwave observations obtained regularly by low-Earth-orbiting satellites usually provide global precipitation measurements with high temporal resolution and low spatial resolution [11]. Since the launch of the Tropical Rainfall Measuring Mission (TRMM) satellite in 1997, satellite remote-sensing-based precipitation products have matured [2]. The TRMM multisatellite precipitation analysis (TMPA) product [7] and its global successor, the integrated multisatellite retrievals for global precipitation measurement (GPM) product, have been used widely in many areas owing to their reasonably high accuracy and general applicability [12–15]. The GPM product provides rainfall estimates from merged multisensor information [7]. Satellite-based rainfall products still have various sources of inherent systematic biases and random errors [12,16–23]. Errors inherent in satellite rainfall include measurement errors and model errors. For example, the magnitude of the sampling error depends on both the type/rate of precipitation and the orbit/swath of the satellite [24]. There are also certain other problems related to estimation of light rainfall and attributable to seasonally dependent biases.

To produce more accurate rainfall products, incorporating multiple ground-based and satellite-based precipitation retrievals are needed. Datasets such as the fifth generation European Centre For Medium-Range Weather Forecasts reanalysis (ERA5) [25,26] represent notable alternatives to rainfall products produced from ground-based or remote-sensing-derived observations [27–29]. In recent years, attention has focused increasingly on merging satellite and reanalysis datasets, such as data assimilation and data fusion. For example, the multi-source weighted-ensemble precipitation (MSWEP) dataset, which has 3 h temporal resolution, takes advantage of satellite-, ground-, and reanalysis-based data to estimate rainfall [28]. The Climate Hazards Group Infrared Precipitation with Station (CHIRPS) data rainfall product merges ground-based and infrared cold cloud duration data [30]. However, these datasets depend on the availability of gauge-based measurements.

The SM2RAIN model towards addressing these problems by inferring or correcting rainfall estimation over land using SM observations from satellites or gauges [31]. This method provides accumulated rainfall estimates instead of the instantaneous rate [31,32]. Most methods based on this approach share the same limitations, linked to the limits of measuring SM from space: rainfall estimated only over land, low accuracy in presence of dense vegetation or complex topography and difficulties in estimating rainfall in case of soil saturation. Currently, SM2RAIN has been applied to different satellite SM products over different regions worldwide with satisfying results. SM2RAIN is able to take advantage of SM information because the difference in SM between specific time intervals can be associated directly with rainfall [33,34]. However, multiple SM datasets are rarely used for rainfall estimation.

Generally, a merged dataset usually outperforms a single dataset. In recent decades, many machine learning methods have been developed for data fusion and forecasting purposes, e.g., support vector machine, random forest (RF), long short-term memory (LSTM), and convolutional neural network (CNN) methods. For example, multiple global precipitation datasets can be merged to improve the spatiotemporal rainfall characterization through machine learning algorithms [35]. RF is a supervised machine learning algorithm and first presented by Breiman [36]. The long short-term memory (LSTM) model (which represents the state-of-the-art recurrent cell in many fields) was first proposed in 1997 [37–43]. CNNs were originally designed to resolve image classification problems and they have been applied effectively for remote-sensing-based image classification [44–48]. Recently, CNNs have been successfully used for soil moisture estimations based on Sentinel multi-source data [49]. Integrating multiple machine learning methods for rainfall retrieving gains rare attention.

The traditional statistical analysis of standard deviation, coefficient of variation and root mean square error of SM can represent the spatial and temporal variability of SM. The spatial and temporal variability of SM is largely affected by soil texture and land cover types [50,51]. Therefore, the rainfall retrieval from SM may also be affected by the soil types and land cover characteristics.

Based on the research gaps discussed above, the objectives of the study were to improve the accuracy of rainfall estimation based on SM2RAIN through merging multiple soil moisture (SM) datasets with machine learning methods. This study will propose an integration framework, which consists of multiple machine learning methods, to use satellite (SMOS, SMAP, and ASCAT) and ground-based soil moisture (SM) observations to derive a precipitation product. The remainder of this paper is structured as follows. Section 2 describes the geographic area, reanalysis data, and remote-sensing-derived and ground-based SM and precipitation data considered in the study. Section 3 presents the methodology and statistical metrics. In Section 4, the results of the assessment of the SM and rainfall datasets are presented, and they are discussed in Section 5. Finally, the derived conclusions are presented in Section 6.

2. Data

2.1. Study Area

This study considered the Babaohe River Basin, which is located in the Upper Heihe River Basin in the arid region of Northwest China. The Babaohe River originates in the Qilian Mountains and flows across Qilian County in Qinghai Province. The area and the elevation of the Babaohe River Basin are 2456 km^2 and 2678–4883 m above sea level, respectively (Figure 1a). The most important water supplies in this region are snow and ice melt. Mean annual precipitation is approximately 400 mm, whereas mean annual evaporation is approximately 1500 mm. There are four dominant land use types (i.e., alpine meadow, subalpine shrub, desert steppe, and alpine steppe) and are two principal soil types (i.e., alpine meadow soil and chestnut soil). The spatial distributions of the land use type and the surface soil types maps are shown in Figure 1b,c.

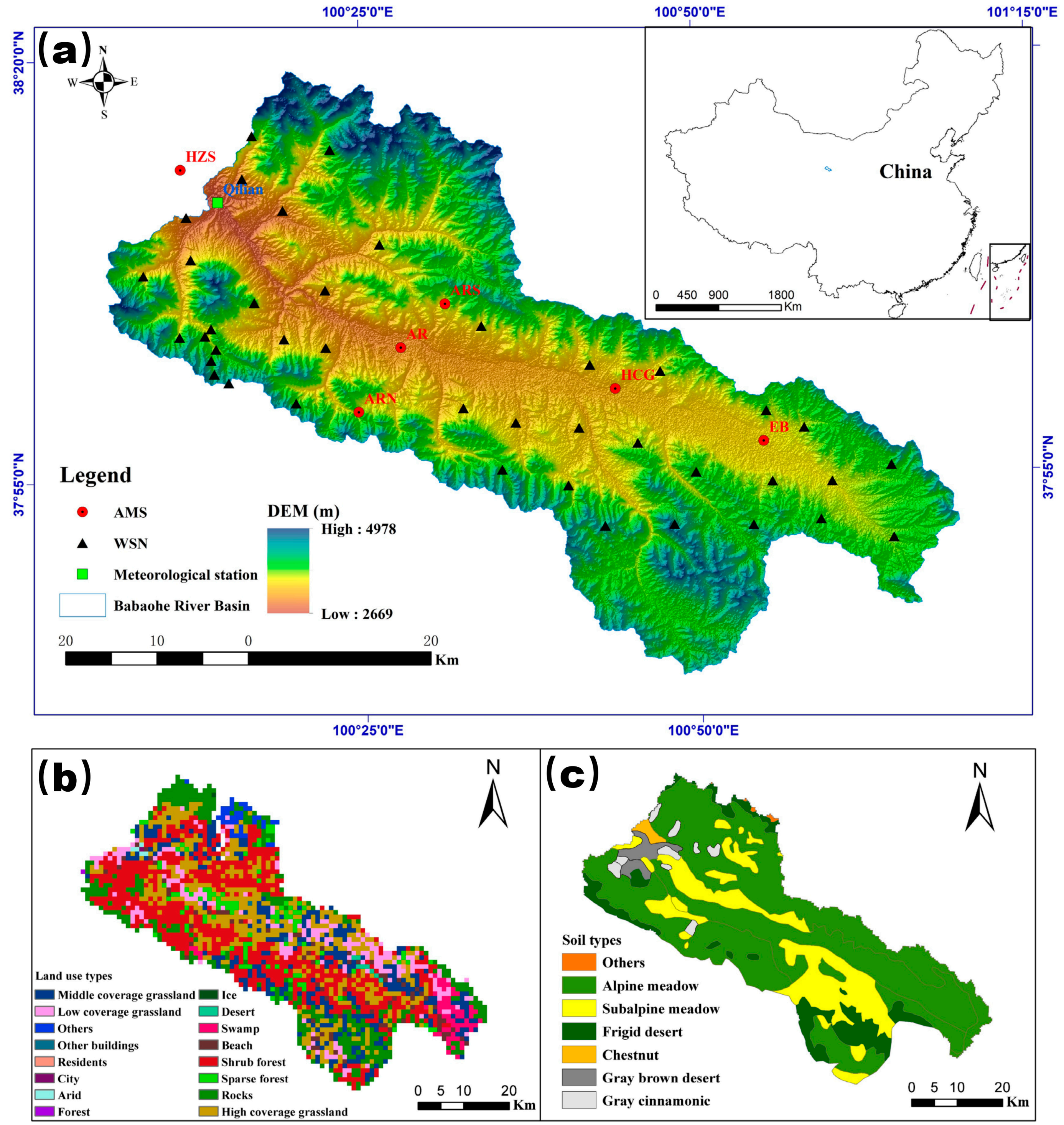

Figure 1. (**a**) Location of the Babaohe River Basin (WSN: wireless sensor network; AMS: automatic meteorological station, including HuangZangShi (HZS), ARou (AR), north slope of ARou (ARN), south slope of ARou (ARS), HuangChaoGou (HCG), and EBo (EB); DEM: digital elevation model). (**b**) Land use types and (**c**) soil types in the Babaohe River Basin.

2.2. China Meteorological Data Service Centre (CMD) Precipitation Dataset

The precipitation data used most widely in China are those recorded at national weather stations. However, some raster datasets derived from these weather station observations are also available, e.g., the precipitation dataset from the China Meteorological

Data Service Centre (CMD). This rainfall dataset is generated from daily rainfall data observed at ground stations, which are interpolated to a 0.5° grid through a bilinear method based on GTOPO30 digital elevation data (0.05° × 0.05°) using ANUSPLIN VERSION 3.2 software [9]. The density of the national weather stations is nearly 700. The elevation data (GTOPO30) was used to regenerate DEM of China to decrease the effect of topography on interpolation of rainfall. The RMSE and R between CMD and stations are 0.49 mm and 0.93, respectively. There are two national meteorological stations (Qilian and Minle) located near the Babaohe River Basin (Figure 1). The CMD dataset is used as reference for calibration of the parameters of the SM2RAIN algorithm (i.e., the soil moisture data retrieval algorithm).

2.3. Remote Sensing Soil Moisture Products

2.3.1. Advanced Scatterometer Data

The Metop-A, Metop-B, and Metop-C satellites acquire Advanced Scatterometer (ASCAT) SM data in the C band. The ASCAT SM product has a spatial resolution of 25 km and daily temporal resolution [52]. For use in this study, ASCAT SM data were obtained from the EUMETSAT project. This product, which has 25 km spatial resolution, has been distributed since 2014. For consistency with other satellite SM and rainfall products, the nearest-neighbor interpolation method was used to reproject the ASCAT dataset to a 0.25° grid. The ASCAT data used in this study (2014–2015) have 0.25° spatial resolution and daily temporal resolution.

2.3.2. Soil Moisture and Ocean Salinity Data

The Soil Moisture and Ocean Salinity (SMOS) data, which are obtained using an L band radiometer, have a spatial resolution of 30–50 km and 3D temporal resolution [53]. In this study, we used Level 3 products distributed by the Centre Aval de Traitement des Données SMOS. These products, which have 25 km spatial resolution, are generated using SMOS L1B products from the European Space Agency (ESA), and are distributed in the most widely used raster format (i.e., NetCDF). The SMOS products are daily products of SM and they contain filtered data. The retrievals are based on a multiorbit retrieval algorithm. Ice and snow cover can be determined and added to the flags. We processed the ascending and descending orbit datasets separately. This product has been produced since 2010 with 0.25° spatial resolution and daily temporal resolution. The SMOS data used in this study (2014–2015) have 0.25° spatial resolution and daily temporal resolution.

2.3.3. Soil Moisture Active and Passive Data

The Soil Moisture Active and Passive (SMAP) SM data, which are obtained by an L-band (1.4 GHz) radiometer, have 36 km spatial resolution and 1–2-d temporal resolution [54,55]. In this study, Level 3 SM retrievals were used. This product is generated from SMAP Level 2 SM data, which are derived from SMAP Level 1C interpolated brightness temperatures. The SMAP data used in this study (2015) have 0.25° spatial resolution and daily temporal resolution.

2.4. Remote Sensing Rainfall Data

The near real-time TMPA rainfall dataset (3B42-RT) has 0.25° spatial resolution and 3 h temporal resolution (accumulated to daily resolution) [7]. Although the grid extends from 60°S to 60°N, high latitude (beyond 50°S/N) near real-time retrievals are considered very unreliable and thus are masked from the daily accumulations. The start and end times for every daily granule are listed in the file's global attributes and they are taken correspondingly from the first and last 3 h granules participating in the aggregation. Therefore, the time period covered by one daily granule amounts to 24 h, which can be inspected in the file's global attributes. Counts of valid retrievals for specific days are provided for every variable, making it possible to compute conditional and unconditional mean precipitation for grid cells where fewer than eight retrievals are available for the day. The TMPA product is available for the period 2000–2020 with 0.25° spatial resolution and daily temporal resolution. The latency of this product is approximately 7 h. The TMPA data used in this study (2014–2015) have 0.25° spatial resolution and daily temporal resolution.

2.5. China Meteorological Forcing Dataset

The China Meteorological Forcing Dataset (CMFD) is a high spatiotemporal resolution gridded near-surface meteorological dataset developed specifically for studies of land surface processes in China [56]. The dataset was produced through fusion of remote sensing products, reanalysis datasets, and ground-based observations at national weather stations [57]. The dataset covers the period 1979–2018 with 3 h temporal resolution and 0.1° spatial resolution. The CMFD provides seven near-surface meteorological elements: 2 m air temperature, surface pressure, specific humidity, 10 m wind speed, downward shortwave radiation, downward longwave radiation, and precipitation rate. For consistency with the other satellite SM and rainfall products used in this study, the nearest-neighbor interpolation method was used to resample the CMFD to 0.25° spatial resolution and daily temporal resolution. The CMFD rainfall data used in this study (2014–2015) have 0.25° spatial resolution and daily temporal resolution.

2.6. GPM-SM2RAIN Dataset

The rainfall dataset (GPM-SM2RAIN) was based on the integration of IMERG-LR of the Global Precipitation Measurement Mission (GPM) with SM2RAIN-based rainfall estimates derived from ASCAT H113 H-SAF, SMOS L3 and SMAP L3 soil moisture products [58]. The dataset is available ±60° (2007–2018) (available at https://zenodo.org/record/3854817 (accessed on 1 October 2022)). The GPM-SM2RAIN we use in this study (2014–2015) was projected to 0.25° spatial resolution and one day temporal resolution.

2.7. Climate Hazards Group InfraRed Precipitation with Station Data (CHIRPS) Dataset

Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) is a 35+ year quasi-global rainfall data set (available at https://www.chc.ucsb.edu/data/chirps (accessed on 1 October 2022)). Spanning 50°S–50°N (and all longitudes) and ranging from 1981 to near-present, CHIRPS incorporates high resolution satellite imagery, and in situ station data to create gridded rainfall time series for trend analysis and seasonal drought monitoring. The hourly CHIRPS we use in this study (2014–2015) was resampled to one day temporal resolution.

2.8. ERA5-Land (ERA5) Dataset

ERA5-Land is a replay of the land component of the ERA5 climate reanalysis, forced by meteorological fields from ERA5 (available at https://www.ecmwf.int/en/era5-land (accessed on 1 October 2022)). ERA5-Land is produced under a single simulation, without coupling to the atmospheric module of the ECMWF's Integrated Forecasting System (IFS) or to the ocean wave model of the IFS. It runs without data assimilation, making it computationally affordable for relatively quick updates. The core of ERA5-Land is the Tiled ECMWF Scheme for Surface Exchanges over Land incorporating land surface hydrology (H-TESSEL). It uses version CY45R1 of the IFS. The ERA5-Land dataset is available for public use for the period from 1950 to 2–3 months before the present with a finer spatial resolution: ~9 km grid spacing. The hourly ERA5 we use in this study (2014–2015) was resampled to one day temporal resolution.

2.9. Multi-Source Weighted-Ensemble Precipitation (MSWEP) Dataset

MSWEP incorporates daily gauge observations and accounts for gauge reporting times to reduce temporal mismatches between satellite-reanalysis estimates and gauge observations (available at http://www.gloh2o.org/mswep/ (accessed on 1 October 2022)). MSWEP tends to exhibit better performance than other precipitation products in both densely gauged and ungauged regions. MSWEP is a global precipitation product with a

3-hourly 0.1° resolution available from 1979 to ~3 h from real-time. The 3-hourly MSWEP we use in this study (2014–2015) was resampled to one day temporal resolution.

2.10. Ground Soil Moisture and Rainfall Data

The ground-based SM and rainfall data were measured by automatic meteorological stations (AMSs) and the Heihe Watershed Allied Telemetry Experimental Research ecohydrological wireless sensor network (WSN) [59]. There are 6 AMS sites and 43 WSN sites within the study area (Figure 1). The ordinary AMS measures radiation, precipitation, air pressure, wind speed and direction, air temperature, humidity, SM, and temperature profiles, and soil heat flux (Figure 2), while hydrological, meteorological, and ecological observations can be acquired by the ecohydrological WSN. The temporal resolution of both the SM data and the rainfall data is minutes.

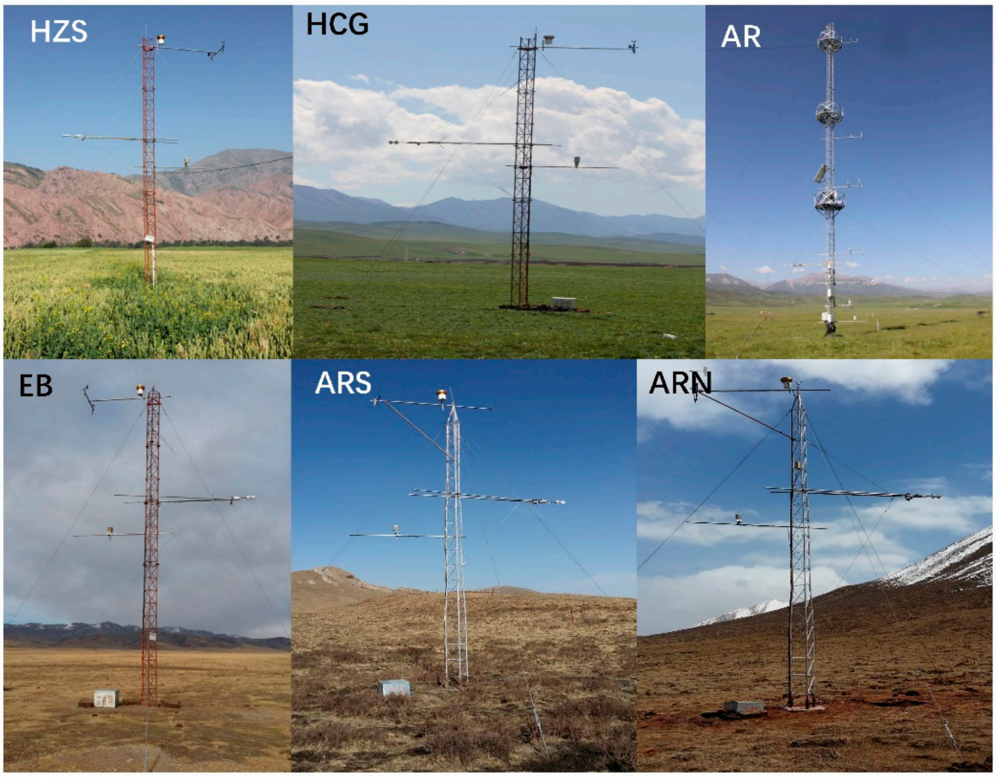

Figure 2. Automatic meteorological stations in the Babaohe River Basin.

3. Methodology

A new framework was proposed in this study to retrieve rainfall from satellite-derived and ground-based SM. First, satellite SM (SMOS, SMAP and ASCAT) and ground SM were trained in RF, LSTM and CNN machine learning models. The trained model (RF, LSTM and CNN) was then used to generate SM (RF-SM, LSTM-SM and CNN-SM) based on satellite SM datasets. The Bayesian model averaging method was used to merge RF-SM, LSTM-SM and CNN-SM based on each data error. The original TMPA was corrected by weather station data first using an empirical bias correction equation. Then, the merged SM and the corrected TMPA rainfall datasets were used to calibrate SM2RAIN model parameters. The AMS and GPM-SM2RAIN rainfall datasets were used to assess SM2R dataset. Finally, SM2R, CMD and CMFD were used for TC analysis (Figure 3).

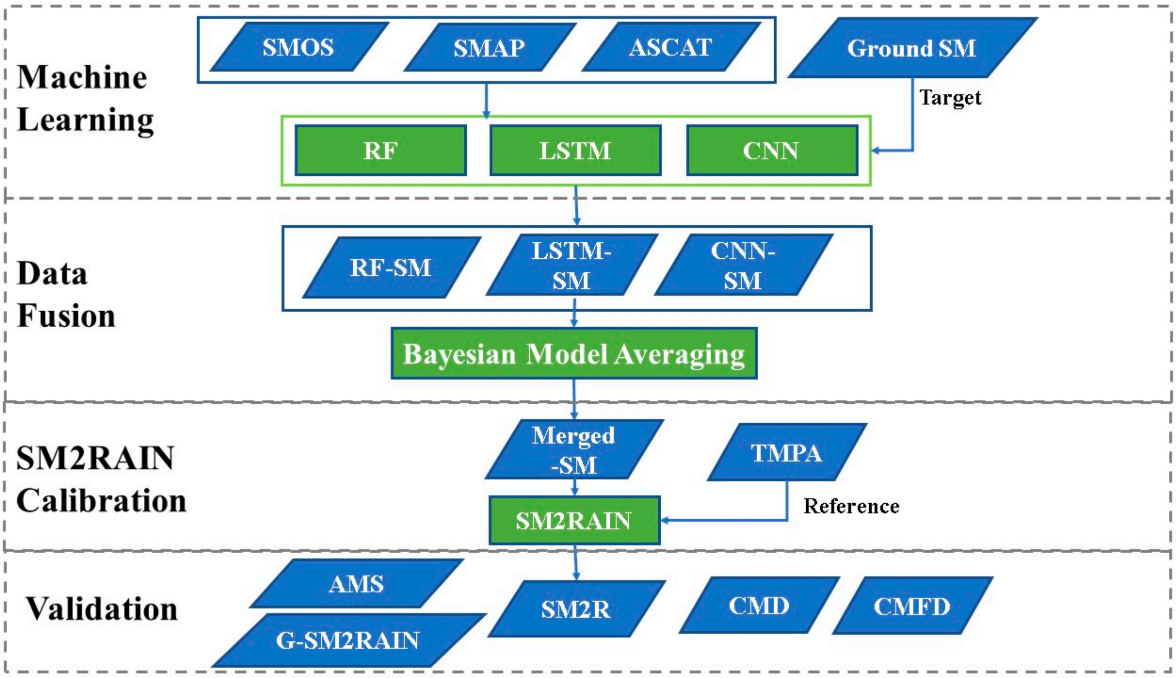

Figure 3. Flowchart of the production of SM2R from ground-based and satellite-derived (SMOS, SMAP, and ASCAT) SM with SM2RAIN and machine learning methods (random forest (RF), long short-term memory (LSTM), and convolutional neural network (CNN)).

3.1. Data Processing

Because SM and rainfall products can present different spatiotemporal resolutions, we resampled the SM datasets (SMOS, SMAP, and ASCAT) and rainfall datasets of the Upper Heihe River Basin to the same spatial and temporal resolutions (i.e., 0.25° × 0.25° and daily, respectively) using the nearest-neighbor interpolation method. Additionally, the CMD rainfall data, characterized by 0.5° resolution, were downscaled to 0.25° resolution using a bilinear interpolation method. Rainfall accumulations were accumulated to the daily scale. A single empirical bias correction equation was used to correct TMPA, that was uniform in space and time, and the details of the bias correction equation can be referred to Vernimmen et al. [60]. Ground SM and rainfall data were separately averaged and accumulated over the period 00:00–23:59 UTC.

3.2. SM2RAIN Model

SM2RAIN is an algorithm used to estimate rainfall from SM observations [31]. This model is a type of inverted single-layer soil water balance model that uses SM observations at current and previous time steps as input. As the evapotranspiration and surface runoff contributions are negligible during a rainfall event, the water balance algorithm can be expressed as follows:

$$P_{sm2rain}(t) = Z^* \frac{ds(t)}{dt} + R(t) + E(t) + G(t), \quad (1)$$

$$P_{sm2rain}(t) = Z^* \frac{ds(t)}{dt} + as(t)^b, \quad (2)$$

where $P_{sm2rain}$ is the estimated rainfall, $R(t)$ is the surface runoff, $E(t)$ is the evapotranspiration, $G(t)$ is the drainage rate, s is the relative saturation of the soil (value: 0–1), t is time, Z^* is the soil water capacity (equal to the soil layer depth multiplied by porosity), and a and b are two parameters that illustrate the nonlinearity between the loss rate (including drainage and evapotranspiration components) and soil saturation. In Equation (1), when the estimates of rainfall are below zero, rainfall values will be zero.

The SM2RAIN parameters (a, b and Z^*) can be obtained by calibration on the basis of the differences between two observed consecutive SM and rainfall data. We used Equation (2) to estimate the rainfall within the time between the two observations [31,61]. When snowfall occurred, the TMPA value was used.

3.3. Model Calibration

All the parameters described in Section 3.2 were calibrated using the rainfall dataset (corrected TMPA). In this process, the daily root mean square error (RMSE) between the SM2RAIN rainfall (SM2R) estimated from the merged SM product and ground rainfall (AMS) during 2014–2015 was minimized. However, because the algorithm of RMSE calibration is based on minimizing variance, which is subject to conditional biases, it might reduce the temporal variability of the estimated rainfall; consequently, the extreme values of precipitation might be underestimated. To overcome this problem, other metrics such as the Kling–Gupta Efficiency index [62] could be applied to consider the performance of extreme value estimation between observations and estimations. Nevertheless, to ensure homogeneity between the model calibration and assessments of rainfall and SM products, we decided to use RMSE. The product obtained from SM2RAIN is referred to as SM2R.

3.4. Machine Learning Methods

3.4.1. Random Forest

The RF is a collection of multiple decision tree models (i.e., a type of flowchart that shows a clear pathway to a decision). The RF is a supervised machine learning algorithm first presented by Breiman [36]. The RF algorithm generates a set of decision trees from a few randomly nominated subsets of a training set and selects predictions from each tree. Finally, the RF algorithm chooses the best solution by means of voting (Figure 4).

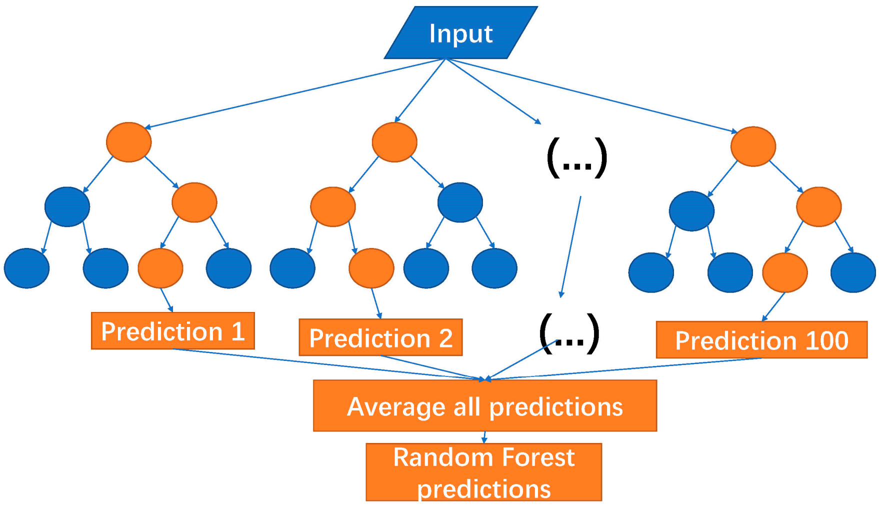

Figure 4. Flowchart of data fusion based on random forest.

When there is a nonlinear relationship between the input values and the target values, decision trees are suitable for determining the relationship between them. The RF algorithm is a type of conception of an ensemble of decision tress [36].

For the RF model used in this study, the generation of the merged soil moisture data (RF-SM) can be divided into two parts. First, SMOS and ASCAT were trained and validated in RF data fusion from January 2014 to March 2015; second, the SMAP, SMOS and ASCAT were trained and validated in RF data fusion from April to December 2015. For each part, multiple (satellite and ground-based) SM time series at each point located within the Upper Heihe River Basin were randomly split into a training dataset (80%) and validation datasets (20%). In training the RF model, the maximum depth of the RF was set to 1–20 to reduce the phenomenon of overfitting. The optimal maximum depth (8) of the RF was determined by comparing the Pearson correlation coefficient (R) value of the training and validation dataset models. Finally, the trained RF was used to predict a new SM dataset that was based on the multiple SM datasets.

3.4.2. Long Short-Term Memory

The LSTM model, which represents the state-of-the-art recurrent cell in many fields, was first proposed in 1997 [37].

LSTM includes two important features. The first feature is an explicit state, which is a separate set of variables that store the elements required to construct long- and short-term dependencies, including the current state. The second feature is the presence of gates. A gate can be considered as an element that can control the amount of information flowing through it. In LSTM, gates are realized using sigmoid functions and their activation is based on hyperbolic tangents. The structural diagram of an LSTM cell is shown in Figure 5.

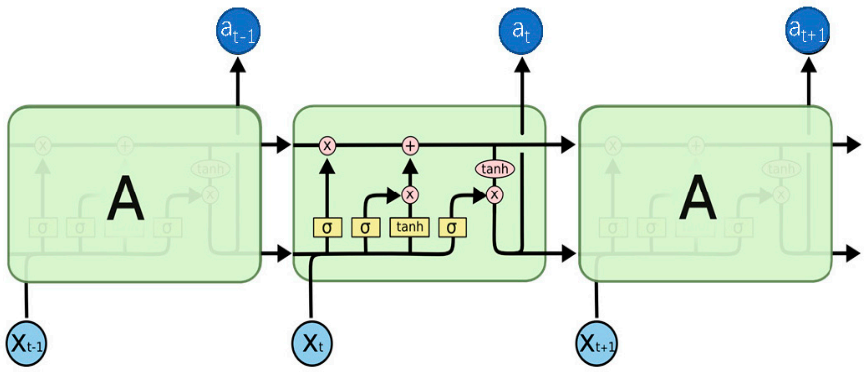

Figure 5. Structural diagram of a long short-term memory cell.

The functions f, i, and o represent the cores of the forget, input, and output gates, respectively. The first element is the memory state, which is responsible for the dependencies and for the actual output. The equations can be expressed as follows:

$$i = \sigma\left(x_t U^i + a_{t-1} W^i\right), \quad (3)$$

$$f = \sigma\left(x_t U^f + a_{t-1} W^f\right), \quad (4)$$

$$o = \sigma(x_t U^o + a_{t-1} W^o), \quad (5)$$

$$g = \tanh(x_t U^g + a_{t-1} W^g), \quad (6)$$

$$c_t = c_{t-1} \circ f + g \circ i, \quad (7)$$

$$a_t = \tanh(c_t) \circ o, \quad (8)$$

$$y = softmax(V a_t), \quad (9)$$

where i is the input gate that is responsible for how much new information is let into the memory cell, f is the forget gate that is responsible for the information that should be thrown away from the memory cell, o is the output gate that controls how much information is passed on to the next time step, g is a self-recurrent connection that is equivalent to a standard recurrent neural network, c_t is the internal memory of the memory cell, s_t is a hidden state, y is the final output, and

$$softmax(b_i) = \frac{\exp(b_i)}{\sum_{j' \in \{1,2,\ldots K\}} \exp(b_j)}. \quad (10)$$

The state is based on the following:

- A dynamic balance between the previous experience and its reevaluation according to a new experience (modulated by the forget gate);
- The semantic effect of the current input (modulated by the input gate) and the potential additive activation.

In this method, the generation of the merged soil moisture data (RF-SM) can be divided into two parts. First, SMOS and ASCAT were trained and validated from Jan 2014 to Mar 2015; second, the SMAP, SMOS and ASCAT were trained and validated from April to December 2015. For each part, multiple (satellite and ground-based) SM time series at each point located within the Upper Heihe River Basin were considered as input training and validation datasets for the LSTM. The network was repeatedly trained by varying the number of hidden layers from 1–10 and increasing the number of neurons in each layer from 1–128 with an increment of 8. Then, the optimal LSTM network structure (6 layers and 68 neurons) was selected on the basis of SM. Finally, the trained LSTM model was used to predict a new SM dataset that was based on the multiple SM datasets.

3.4.3. Convolutional Neural Network

A CNN is a distinct type of deep network that reflects the local topology of data. A CNN comprises a fully connected network and a constrained network that includes convolutional and pooling layers. The convolution and pooling operations are used in the constrained network to realize the local receptive field and parameter reduction.

CNNs were originally designed to resolve image classification problems and they have been applied effectively for remote-sensing-based image classification [44,45]. This study used the deep learning method to learn the direct mapping between the ground-based SM and the satellite-derived SM. The CNN was used to train the ground-based SM and the corresponding satellite-derived SM to encode the mapping, and the fusion rules of the merged SM were generated through CNN model learning. These rules determine the ground and satellite information of the merged SM. In recent decades, experiments have demonstrated that CNN-based deep learning methods are reliable for predicting SM using multisource data [46–48,63]. Recently, CNNs have been used successfully for SM estimations based on Sentinel multisource data [49].

In this method, the CNN was applied to prediction of SM at a given position and time using satellite, and ground-based SM time series split into training (80%) input and validation (20%) input. The generation of the merged soil moisture data (CNN-SM) can be divided into two parts. First, SMOS and ASCAT were trained and validated from Jan 2014 to Mar 2015; second, the SMAP, SMOS and ASCAT were trained and validated in from April to December 2015. The input time series at each ground-based SM point comprised observations with the same time step. A 1D CNN model can learn a mapping from the input sequences to an output value. CNNs can support parallel input SM time series as separate channels, such as the red, green, and blue components of an image. Therefore, the data should be split into samples that maintain the order of the observations across the input sequences (e.g., SMOS, SMAP, and ASCAT). Then, the time series can be transformed into input/output samples to train the model. Each input series can be handled by a separate CNN, and the output of each of these CNNs can be combined before producing the prediction as the output sequence, which can be referred to as a multilayer perceptron.

We defined this type of CNN model in Keras (an artificial neural network Python package) using the Keras functional application programming interface. First, we defined the first input model as a 1D CNN with an input layer that expects vectors with total steps and one feature. Second, we defined multiple 1D CNN models in the same way. Third, we defined all the input models, and we merged the output from each model into one vector that could be interpreted before producing the prediction as the output sequence. Fourth, we fitted the CNN model defined above using the input SMOS, SMAP, ASCAT, and ground-based SM time series. Finally, we ran the model to predict the SM time series for each point.

The workflow of the CNN model developed for producing SM dataset fusion (prediction) is shown in Figure 6.

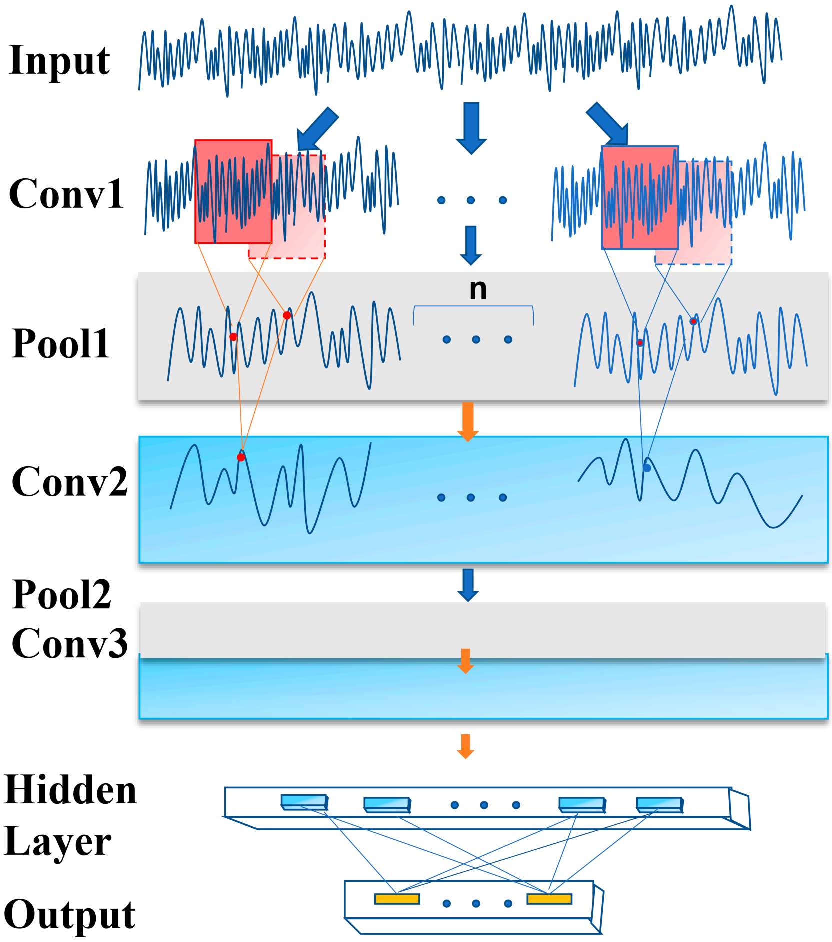

Figure 6. Flowchart of convolutional neural network for merging multiple soil moisture time series.

3.5. Classical Validation

Three widely used metrics were adopted to evaluate the performance of the SMOS, SMAP, ASCAT, and merged SM products and the CMD, TMPA, CHIRPS, ERA5, MSWEP and SM2R rainfall products: the mean relative error (BIAS, in mm), R, and RMSE (in mm), which were computed for each grid and each time on the basis of the accumulated rainfall throughout the entire investigated period. The continuous metrics were adopted for validating both daily and monthly SM and rainfall data at given times and locations. The performance metrics were calculated as follows:

The value of R was calculated as follows:

$$\mathrm{R} = \frac{\sum_{t=1}^{M} \left(\overline{Qs_t} - Qs_{t,g}\right)\left(\overline{Qo_t} - Qo_t\right)}{\sqrt{\sum_{t=1}^{M} \left(\overline{Qs_t} - Qs_t\right)^2}\sqrt{\sum_{t=1}^{M} \left(\overline{Qo_t} - Qo_t\right)^2}}. \tag{11}$$

The additive bias was calculated as follows:

$$\mathrm{BIAS} = \frac{\sum_{t=1}^{M} \left(\overline{Qs_t} - Qo_t\right)}{M}. \tag{12}$$

The averaged RMSE was calculated as the temporal and spatial averaged RMSE for SM or rainfall, separately:

$$\mathrm{RMSE} = \sqrt{\frac{1}{MN}\sum_{t=1}^{M}\sum_{g=1}^{N}\left(\overline{Qs_{t,g}} - Qo_{t,g}\right)^2}, \tag{13}$$

where $Qo_{t,g}$ is observed variable value at time t and grid cell g, M is the maximum number of time steps, N is the maximum number of grid cells, and $\overline{Qs_{t,g}}$ is the ensemble mean simulated variable value at time t and grid cell g.

3.6. Triple-Collocation Analysis

In this study, triple-collocation (TC) analysis [64] was used to estimate the errors of the rainfall product (SM2R) derived from SM2RAIN with CMD and CMFD rainfall products. Here, we present a short description of the theory underlying TC. Suppose we have three measurement systems X_i, observing a true variable t that is characterized by an additive error model:

$$X_i = \alpha_i + \beta_i t + \varepsilon_i, \tag{14}$$

where variables X_i (i = 1, 2, 3) are collocated measurement systems linearly related to the true underlying value t with additive random errors ε_i,, while α_i and β_i are the ordinary least square intercepts and slopes separately. Assuming that the errors from the independent sources have zero mean ($E(\varepsilon_i)$ = 0) and are uncorrelated with each other (Cov(ε_i;ε_j) = 0, with $i \neq j$) and with t (Cov(ε_i,t) = 0), the variance of the error of each dataset can be expressed as follows Chen et al. [65]:

$$\sigma_\varepsilon = \begin{bmatrix} \sqrt{Q_{11} - \frac{Q_{12}Q_{13}}{Q_{23}}} \\ \sqrt{Q_{22} - \frac{Q_{12}Q_{23}}{Q_{13}}} \\ \sqrt{Q_{33} - \frac{Q_{12}Q_{23}}{Q_{12}}} \end{bmatrix}, \tag{15}$$

where Q_{ij} = Cov(X_i; X_j) is the covariance within the variables X_i.

3.7. Spatial and Temporal Analysis of SM Products

The temporal analysis of SM was performed based on the annual mean of SM (Mean), standard deviation (Std) and the coefficient of variation (CV) of SM within the entire watershed. The spatial analysis of SM was calculated based on the annual mean of SM, standard deviation and the coefficient of variation of SM in different underlying surface conditions of the watershed. The calculation of CV is as follows:

$$\mathrm{CV} = \frac{\mathrm{Std}}{\mathrm{Mean}} \tag{16}$$

4. Results

The presentation of the results is split into three parts: (1) assessment of the merged SM product with WSN, (2) assessment of the SM2R product with ground-based rainfall datasets, and (3) assessment of the SM2R product through TC with the CMD and CMFD products.

4.1. Assessment of the Merged Soil Moisture Product with WSN

The merged SM product is shown at the monthly time step in Figure 7. During winter or the dry season (January–March), surface SM is distributed reasonably uniformly within the Upper Heihe River Basin. During summer and autumn or the wet season (April–June, August, October, and November), southern parts of the Babaohe River Basin have relatively high SM, reflecting the reasonably high vegetation coverage in the south of the Babaohe River Basin. The daily SM RMSEs for RF, CNN and LSTM is 0.052 m^3/m^3, 0.06 m^3/m^3 and 0.058 m^3/m^3. Based on Bayesian theory, the uncertainties of the models can be decreased by using Bayes model averaging (BMA) [66]. Therefore, the merged SM (SM_m) was generated by using BMA with three SM datasets. The equation of the BMA to generate merged SM with RF-SM (SM_1), LSTM-SM (SM_2), and CNN-SM (SM_3) can be expressed as:

$$SM_m = \sum_{i=1}^{M} w_i (SM_i) \quad (17)$$

where w is the weight which was calculated based on the error distributions. The weight w was calculated as errors of SM_i divided by sum of all the errors of SM.

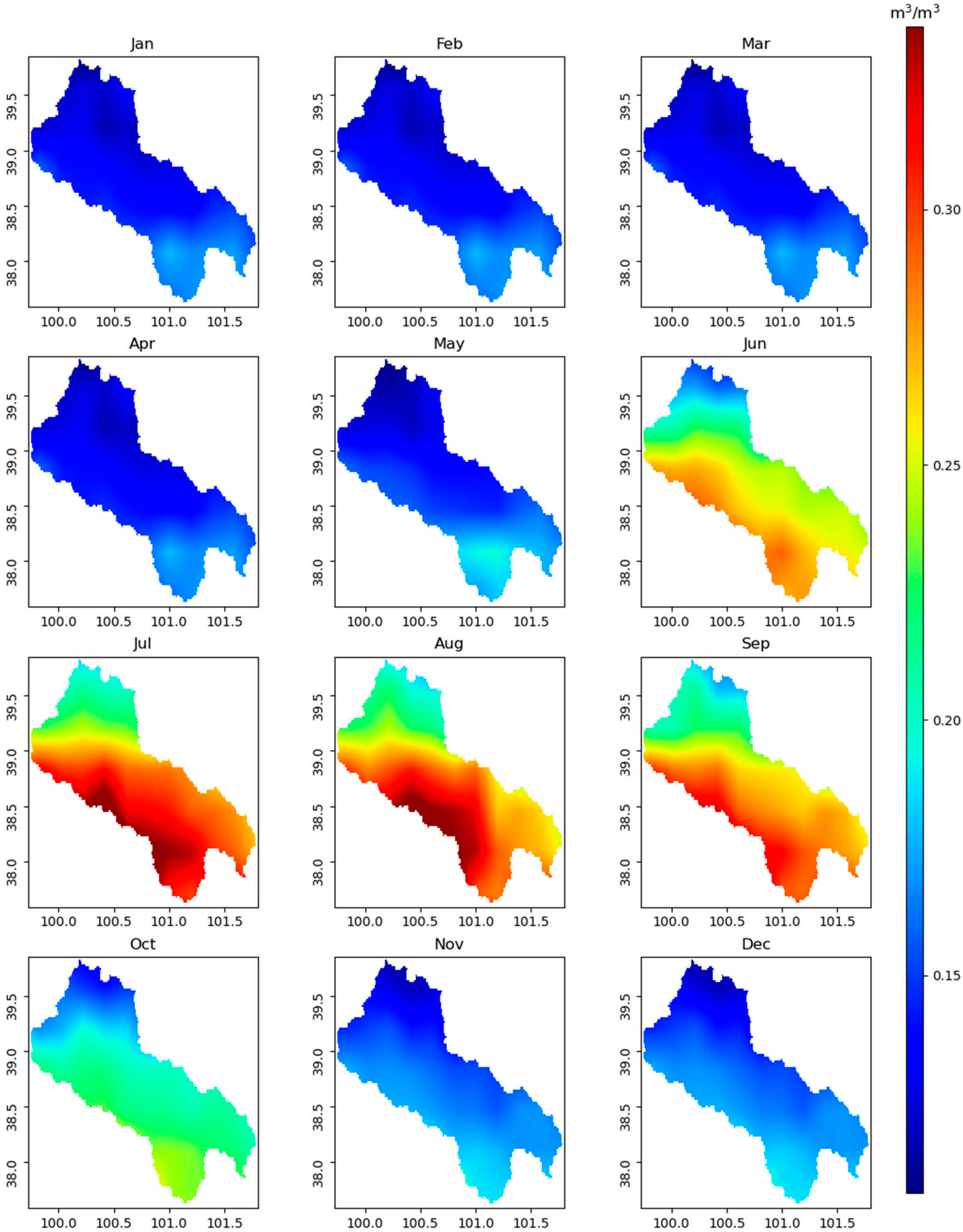

Figure 7. Monthly soil moisture values of merged soil moisture product in 2014.

Figure 8 shows the monthly RMSEs (merged SM product and WSN) of SM at the WSNs during the study period (2014–2015). The overall RMSEs of SM at all WSNs is nearly 0.05 m^3/m^3. The merged SM agrees well with the ground-based SM with an R value of 0.8 and bias of -0.06 m^3/m^3. The results of the merged SM dataset show that it not only retains the spatial distribution of the satellite-derived data but also has greater accuracy than that of a single dataset (compared with SMOS (RMSE = 0.1 m^3/m^3, R = 0.29), SMAP (RMSE = 0.06 m^3/m^3, R = 0.6), and ASCAT (RMSE = 0.06 m^3/m^3, R = 0.42)).

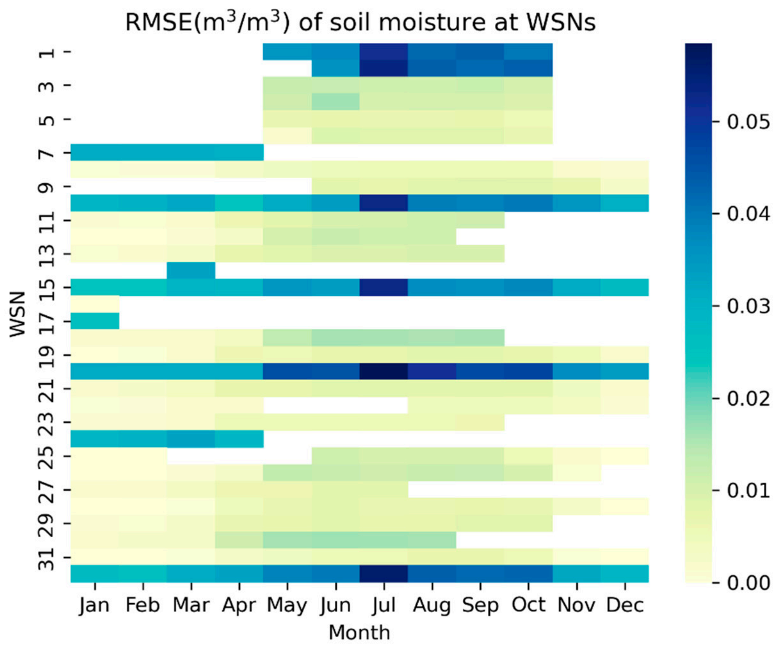

Figure 8. Heat plots of monthly RMSEs of merged soil moisture at the WSNs in 2014.

The merged SM dataset shows substantial errors during the wet period (April–October). Rainfall can dramatically affect the moisture condition of soil and the higher the SM content, the higher the errors in the SM product. Some sites have insufficient data for plotting the RMSEs; therefore, data from only 32 sites considered in this study are plotted. Overall, the above results further demonstrate the effectiveness of the merged SM, which means that the merged SM dataset is suitable for high-precision hydrology and drought monitoring applications.

4.2. Classical Validation of SM2R Using High-Quality Ground-Based Observations

The SM2R dataset at the monthly time step is shown in Figure 9. Rainfall events generally occur in central and northern parts of the Babaohe River Basin. However, the distribution of rainfall is nonuniform and extreme rainfall usually occurs in central or southeastern parts of the Babaohe River Basin, associated with the orographic effects of the Qilian Mountains (elevation in the southeast is slightly higher than in the northwest). Compared with CMFD (Figure 10) and TMPA (Figure 11), SM2R showed similar spatial distributions with CMFD and TMPA (rainfall usually occurs in the southern part of the Babaohe River Basin), but lower rainfall estimates. Additionally, TMPA showed higher estimates and lower accuracy compared with interpolated daily rainfall (such as CMD) [67].

The performance metrics (RMSE, R, and bias) of most stations are acceptable most of the time. Sites ARS, HZS, and HCG have lower R values and higher values of RMSE and bias in comparison with the other sites (Figure 12). There are large errors in the SM2R product at the ARS and HCG sites due to the sparse grassland. During July and August, the RMSEs at site ARS are dramatically higher than those at other sites. Extreme rainfall (storm) usually occurs during July and August (Figure 12). However, the SM2R dataset presents higher R values at the ARS site, which is located on the southern slope of a mountain near the AR superstation (Figure 1). The ARS site is a typical desert steppe station (Figure 2). Therefore, when storm rainfall occurs, the soil will quickly become saturated owing to its fragile surface and thus the low fluctuations will not truly reflect the strength of the rainfall, which plays an important role in rainfall retrieval. Consequently, the SM at sites with this type of soil structure will not reflect the rainfall conditions after a storm rainfall event; the stronger the storm, the greater the errors presented in the rainfall product. In addition to ARS, site HCG is also in a sparse grassland location (Figure 2).

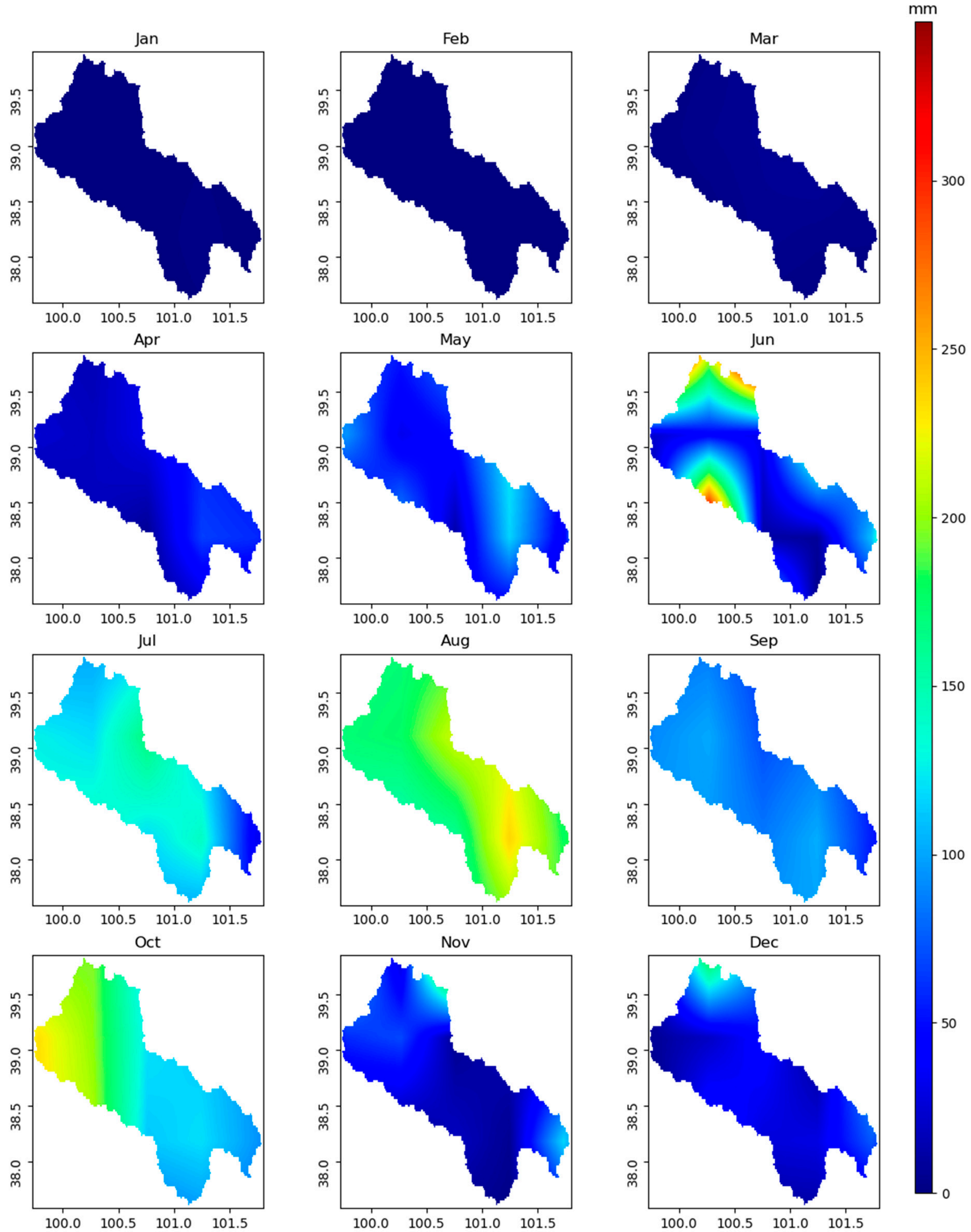

Figure 9. Monthly rainfall estimates of the SM2R product in 2014.

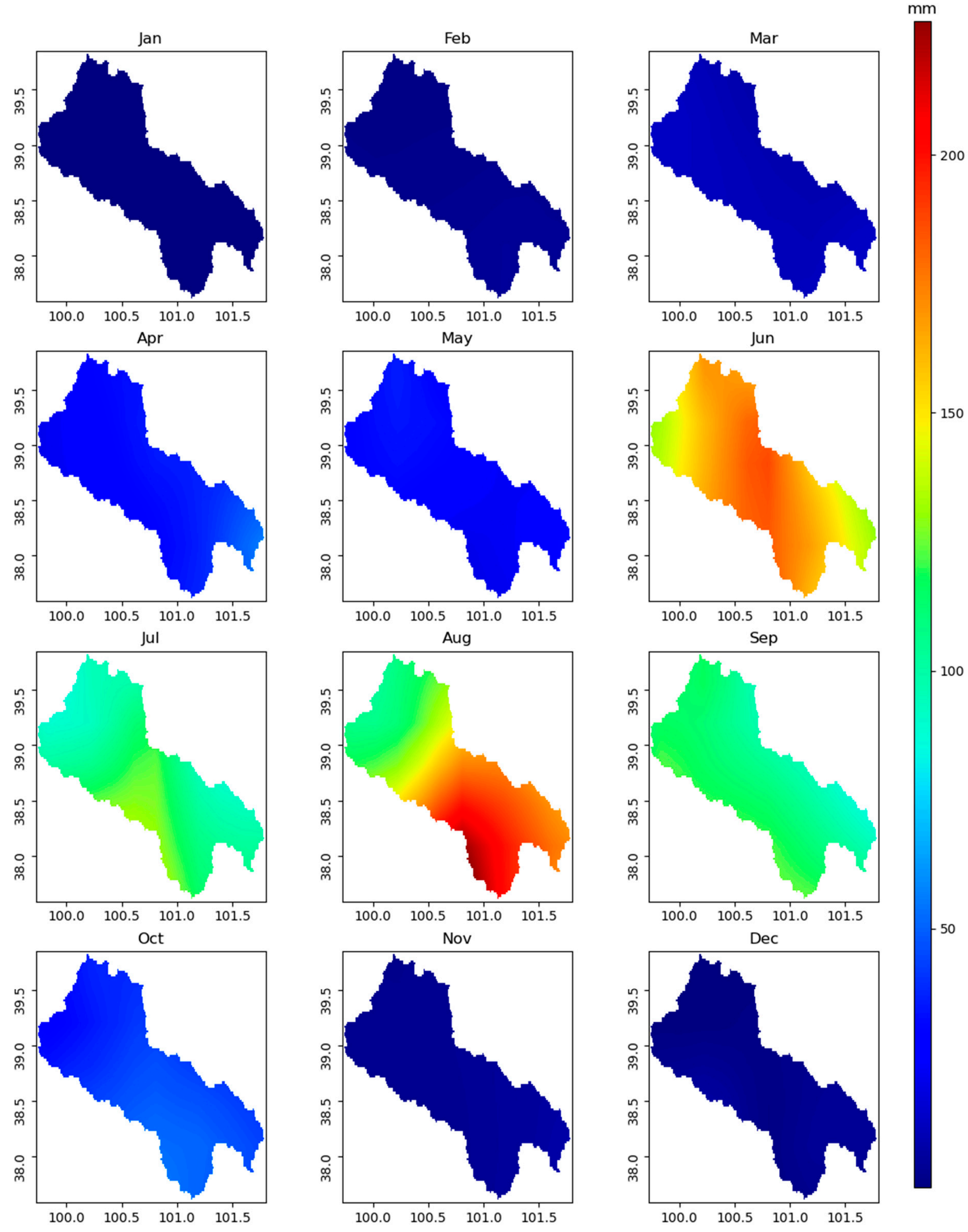

Figure 10. Monthly rainfall estimates of the CMFD product in 2014.

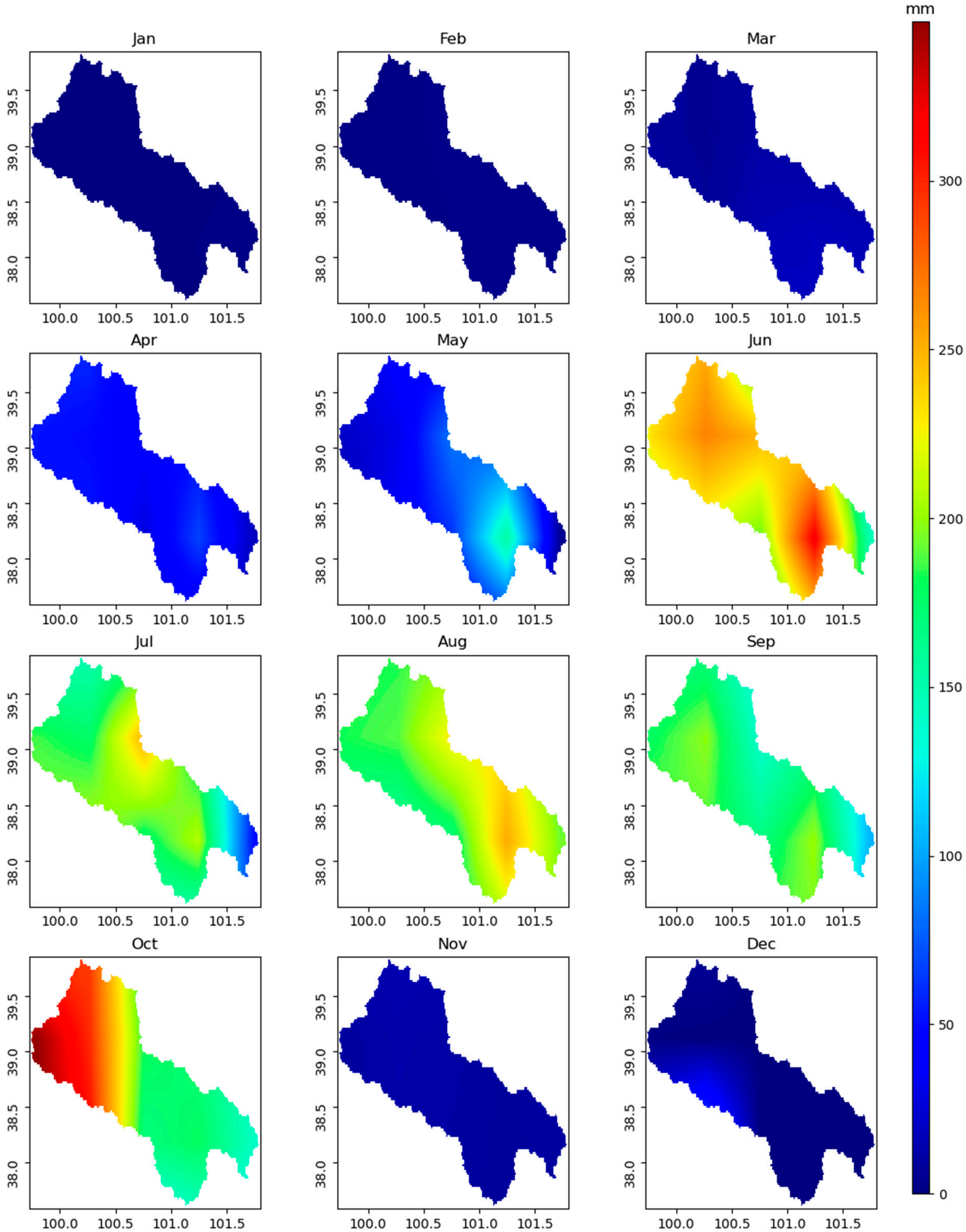

Figure 11. Monthly rainfall estimates of the TMPA product in 2014.

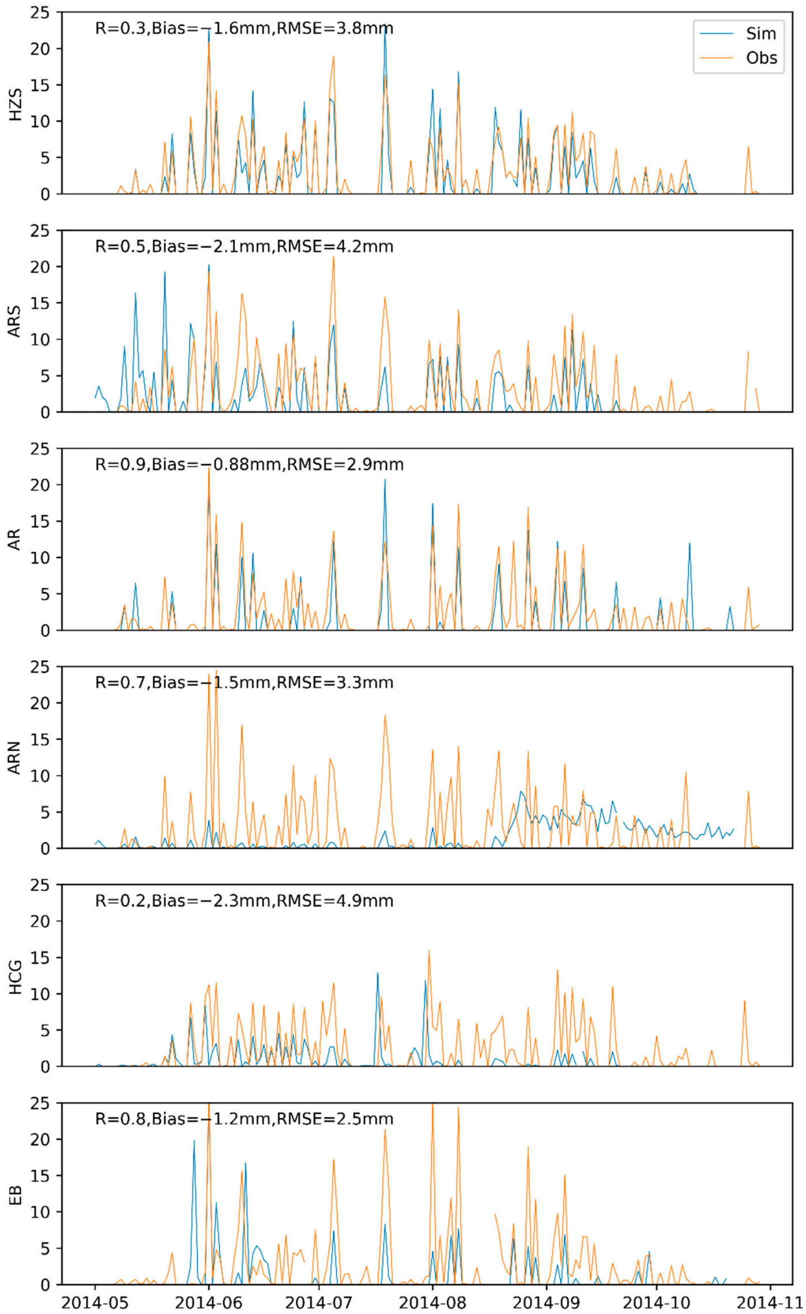

Figure 12. Daily rainfall estimates at the AMSs.

The R value of site HZS is lower (higher) than that of AR, ARN, EB, and ARS (HCG), and the RMSE and bias values of HZS are higher (lower) than those of AR, ARN, EB and ARS (HCG) (Table 1). Site HZS is located in an area of cropland. The jointing period of wheat occurs after April; therefore, during the investigated period, the cropland surface reflected either low or high coverage of vegetation (Figure 2). In comparison with the other sites, the SM of HZS is most affected by human activities (irregular irrigation). Therefore, the performance metrics of site HZS are worse than those of most other sites. The performance metrics of site HCG are the worst among all the investigated sites. The surface of site HCG has the lowest coverage of grass among all the sites, and the evapotranspiration associated with the sparse grass affects the fluctuation of surface SM dramatically. Additionally, root uptake of surface SM under conditions of low grass coverage is also much larger than that under conditions of higher grass coverage. Consequently, SM fluctuations cannot reflect only the rainfall effect, and there will be large errors in the SM2RAIN rainfall estimates for areas with sparse grass coverage, which mostly reflect the effects of evapotranspiration and root uptake. However, evapotranspiration in the mountains of Western China is dramatically high, even during a rainfall event. Therefore, evapotranspiration should be considered in SM2RAIN in areas with high and rapidly changing evapotranspiration.

Table 1. Performance metrics of SM2R/GPM-SM2RAIN/CHIRPS/ERA5/MSWEP rainfall data.

AMS	R	RMSE (mm)	Bias (mm)	Land Cover	Soil Type
AR	0.9/0.62/0.25/0.28/0.27	1.98/2.32/2.62/4.37/2.71	−0.88/−1.1/0.99/2.31/1.01	High Coverage Grassland	alpine meadow soil
ARN	0.7/0.55/0.35/0.7/0.54	3.30/3.63/3.65/4.2/3.61	−1.51/−1.9/1.48/1.93/1.49	High Coverage Grassland	alpine meadow soil
EB	0.84/0.61/0.26/0.32/0.25	2.58/3.5/3.39/4.8/3.23	−1.21/−1.7/1.79/2.1/1.94	Middle Coverage Grassland	alpine meadow soil
HCG	0.17/0.23/0.13/0.64/0.48	4.86/4.38/4.75/5.1/3.71	−2.29/−2.1/2.46/2.39/2.77	Low Coverage Grassland	alpine meadow soil
HZS	0.32/0.26/0.1/0.6/0.5	3.83/4.02/4.34/4.2/2.74	−1.61/−1.8/1.73/2.13/1.66	Cropland	chestnut soil
ARS	0.563/0.45/0.29/0.23/0.3	4.20/4.14/4.32/5.1/4.13	−2.1/−2.5/2.81/2.2/2.24	Low Coverage Grassland	alpine meadow soil

For most of the AMS sites with moderate or high grass coverage, the SM2R product developed in this study has reasonable agreement with the ground-based SM observations, indicating that the integration successfully excludes low-quality SM information.

For different rainfall products, the performance ordered by R (Table 1): SM2R (0.58) > ERA5 (0.46) > GPM-SM2RAIN (0.45) > MSWEP (0.39) > CHIRPS (0.23); the performance ordered by RMSE: MSWEP (3.35 mm) < SM2R (3.45 mm) < GPM-SM2RAIN (3.66 mm) < CHIRPS (3.85 mm) < ERA5 (4.62 mm); the performance ordered by Bias: SM2R (−1.6 mm) < GPM-SM2RAIN (−1.85 mm) < MSWEP (1.852 mm) < CHIRPS (1.88 mm) < ERA5 (2.18 mm). After all, SM2R is best among GPM-SM2RAIN, CHIRPS, ERA5 and MSWEP.

Overall, the above results further demonstrate the effectiveness of the SM2R product and indicate its suitability for operational and near real-time hydrological applications and climate change analyses.

4.3. Validation of SM2R Using Triple-Collocation Analysis

The errors for the SM2R dataset calculated through TC analysis are shown in Figure 13. Areas that present high SM2R uncertainties often reflect mountains with high coverage of vegetation; some residential areas show that the SM2R product has low accuracies.

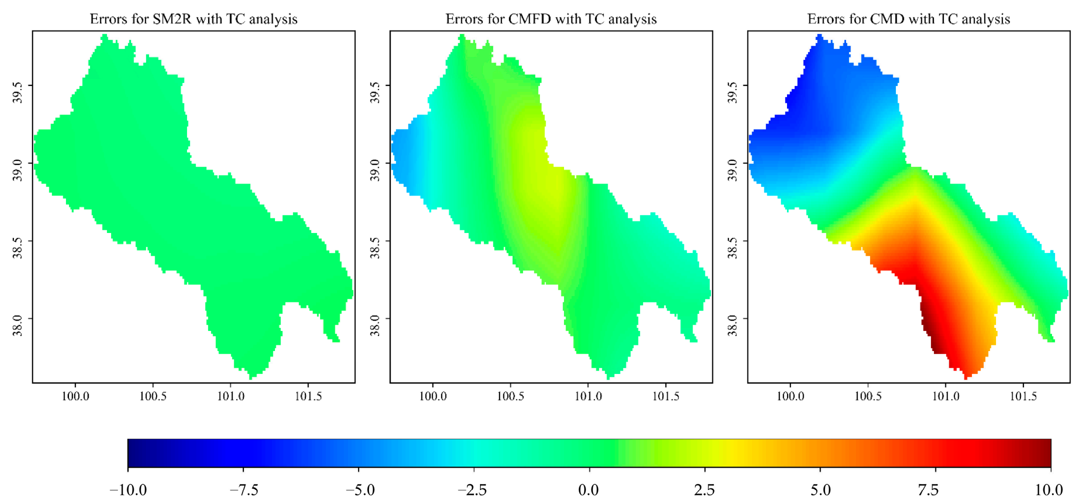

Figure 13. Errors for SM2R, CMFD and CMD rainfall products through TC analysis (Unit: mm).

Assessment of the SM2R, CMFD, and CMD rainfall products over the Upper Heihe River Basin was conducted using TC analysis. Overall, reasonably good performance was obtained by the SM2R rainfall product in data-sparse areas.

4.4. Spatial and Temporal Analysis of SM and SM2R

The temporal variations of SM are illustrated in Table 2. The standard deviation (Std) was highest in autumn, followed by summer, spring and winter. The coefficient of variation (CV) was highest in spring, followed by autumn, summer and winter. In spring (March–May), the SM increased rapidly. As the temperature increased, permafrost melt and rainfall increased, and soil moisture increased. In summer (June–August), the SM was highest. Rainfall frequently occurred and SM fluctuated as the rainfall occurred. The variation was only nearly 0.06 m^3/m^3; therefore, the Std is large, and CV is low. In autumn (September–November), the SM decreased rapidly. Rainfall decreased and the evapotranspiration was still large; therefore, SM decreased. In winter (December–February), the SM was stable and lowest. When the temperature decreased and soil froze, soil moisture became solid and nearly no rainfall compensated the soil moisture, and evapotranspiration was low, the SM was lowest and stable.

Table 2. Statistics of SM data in 2014.

Month	Range (m^3/m^3)	Median (m^3/m^3)	Mean (m^3/m^3)	Std (m^3/m^3)	CV
1	0.036–0.045	0.04	0.04	0.002	0.05
2	0.035–0.052	0.044	0.043	0.005	0.105
3	0.043–0.131	0.083	0.079	0.028	0.349
4	0.104–0.333	0.274	0.241	0.081	0.335
5	0.282–0.371	0.338	0.326	0.029	0.09
6	0.343–0.453	0.384	0.382	0.023	0.061
7	0.309–0.449	0.368	0.372	0.041	0.11
8	0.292–0.389	0.363	0.354	0.024	0.067
9	0.323–0.381	0.358	0.356	0.017	0.049
10	0.232–0.341	0.321	0.308	0.032	0.102
11	0.077–0.261	0.135	0.153	0.053	0.346
12	0.04–0.038	0.05	0.05	0.01	0.158
Annual	0.035–0.453	0.308	0.241	0.135	0.561

Figures 14 and 15 showed the linear regression between coefficient of variation of soil moisture content CV and its mean daily value in different soil and land cover conditions. In all conditions, as the SM increased, CV decreased. In the silty loam and range shrubland conditions, the CV corresponds well with the mean of SM; therefore, the spatial variations of SM can be presented more regularly in the silty loam and range shrubland conditions. Corresponding to the rainfall event, the rainfall product has higher accuracy when the underlying surface consists of silty loam soil or range shrubland (Table 1).

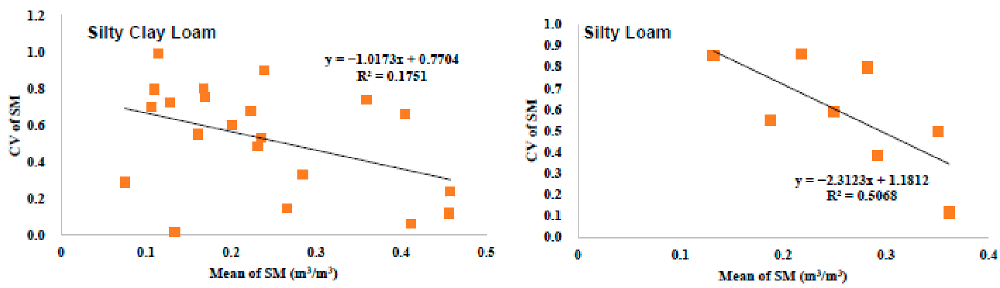

Figure 14. The linear regression between the coefficient of variation of soil moisture content CV and its mean daily value in different soil conditions.

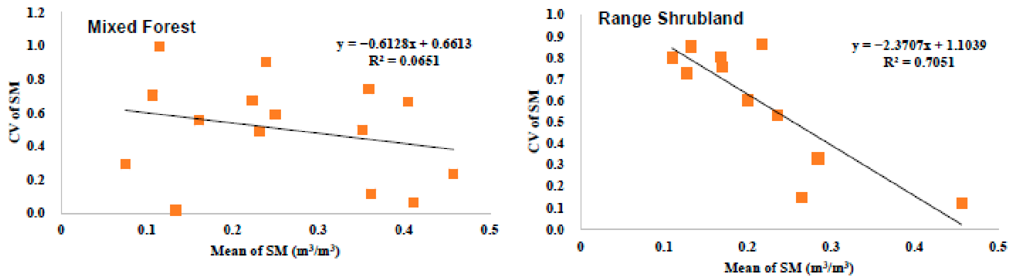

Figure 15. The linear regression between the coefficient of variation of soil moisture content CV and its mean daily value in different land cover types.

5. Discussion

In this study, a daily 0.25° resolution SM and rainfall product was developed by integrating multiple sources of SM and rainfall data, inverted using the water balance model and machine learning methods. The proposed SM and rainfall product can be used for hydrological, meteorological, agricultural and disaster research and applications, such as weather and flood forecasting, water resource management and climate change analysis. The performance of SM2R was assessed by comparative analysis with in situ observations and comparing the results with two state-of-the-art precipitation products namely TMPA and CMD. In general, the overall results (Figure 13) revealed that the SM2R product can effectively capture the precipitation compared with other product. Several pieces of research have been performed in other parts of the world to provide timely and preliminary evaluation reports of newly developed precipitation products [22,23]. They concluded that the accuracy and reliability are mainly dependent on the climatic and geographic condition of that specific region. Hence, this study used different performance metrics to evaluate the accuracy and reliability of a novel product for different land cover and soil conditions (Figures 14 and 15). The spatial variations of SM can be present high linear dependence in the silty loam and range shrubland conditions. Corresponding to the rainfall event, the rainfall product has higher accuracy when the underlying surface consists of silty loam soil or range shrubland (Table 1).

Three machine learning methods, RF, CNN, and LSTM, were used to merge the satellite-derived (SMAP, SMOS, and ASCAT) SM datasets and ground-based SM observations. In extreme (storm) precipitation, all SM datasets have low accuracy. In most microwave SM datasets, when there is heavy rain, the SM value is either invalid or the SM saturation value [68]. Such errors can be diminished through land surface data assimilation based on multiple SM products [4,69,70]. The merged SM dataset is also affected by the characteristics of the underlying surface. Desert steppe areas have low accuracies attributable to the fragile soil texture. Overall, the spatiotemporal resolution and the accuracy of the merged SM dataset were improved in comparison with those of a single SM dataset. More satellite data can also be integrated, such as AMSR2 soil moisture data or GPM rainfall data. More ground-based SM data can be integrated, such as the International Soil Moisture Network (ISMN). More machine learning methods, especially deep learning methods, such as generative adversarial network (GAN), deep belief networks (DBN) and autoencoders (AE) can be integrated into our framework [71].

The SM2R product, generated through merging SM datasets and the TMPA satellite-derived rainfall dataset, performed better than all other rainfall products (i.e., CMFD and CMD). However, its accuracy is dramatically affected by the quality of the SM dataset. The quality of the SM2R rainfall product matched well with the quality of the SM data, although it is also affected by land use and soil texture. Similar conclusions can be drawn regarding SM2R as well as the merged SM product. Owing to the disadvantages of using the RMSE as a metric of SM2RAIN performance, the estimates might contain large biases in extreme rainfall events. Therefore, improved estimates of extreme rainfall might be obtained from SM2RAIN if it were calibrated using the Kling–Gupta efficiency index or other appropriate metrics. A meta-analysis of the effects of grassland degradation on the plant and soil properties of alpine meadows on the Qinghai–Tibetan Plateau revealed that alpine meadow SM decreases significantly ($p < 0.05$) at all depths and for all stages of degradation of vegetation [72]. In comparison with areas with high coverage of grassland, areas with low coverage of grassland have lower SM; consequently, the lower the coverage of grassland, the lower the accuracy of SM2R rainfall data. The SM2R product can be applied to hydrological forecasting in the Babaohe river basin and the method we proposed can be applied for any regions with global available satellite products (SMOS, SMAP, TMPA, etc.) and ground observations (ISMN, etc.).

The coverage of grass affects surface SM through evapotranspiration and root uptake, especially in areas with low coverage. In such areas, SM fluctuations will not truly reflect the rainfall effect, i.e., there will be large errors in the SM2RAIN rainfall estimates that reflect the effects of evapotranspiration and root uptake. There are findings that present the improvement of the rainfall estimates due to the inclusion of the evaporation dataset (ERA5) in the south-western US and central western Australia. However, the evapotranspiration dataset might be not available for the implementation of the processing algorithm in an operational context. Therefore, evapotranspiration should be considered in SM2RAIN in areas with high and rapidly changing evapotranspiration. However, the real-time actual evapotranspiration estimation is still a changeling.

6. Conclusions

Using daily 0.25° resolution SM and rainfall datasets, this study developed a method to integrate multiple sources of SM and rainfall data, inverted using the water balance model and machine learning methods. The SM2R product, which merged multiple SM and rainfall dataset, performed better than all other rainfall products (i.e., CMFD and CMD) in Babaohe river basin. The errors of the SM2R are affected by ground conditions, including land cover and soi types.

Using machine learning methods (RF, CNN, and LSTM), the multiple sources of SM data were used to produce an improved SM product. The errors of the merged SM product depend largely on land use, soil texture, and rainfall intensity. The largest errors occurred in areas of sparse grassland and cropland or during storms. The proposed framework is offline and can be implemented easily and efficiently. Additionally, the proposed framework can also provide single or ensemble forecasting with single or multiple models and multiple datasets. In our next work, we will test our method with more data in this study area and immigrate our framework to other study areas with more satellite data and more machine learning methods integrated, such as generative adversarial network (GAN), deep belief networks (DBN) and autoencoders (AE).

Based on the merged SM products, the SM2R rainfall product derived from SM2RAIN is considered a successful rainfall product with acceptably low spatiotemporal errors. The errors of the SM2R product also depend largely on land-use type and soil properties. The SM2R data have satisfactory quality for grassland areas with moderate vegetation coverage. However, SM retrieval using either satellite or reanalysis technologies remains a problem for farmland areas globally. There are a lot of SM2RAIN-derived rainfall products, such as GPM-SM2RAIN, SM2RAIN-ASCAT and SM2RAIN-CCI. Notwithstanding this, these datasets usually mask out regions characterized by high topographic complexity, frozen soil and high snow probability. Though the SM2R has the same limitations with SM2RAIN—rainfall estimated only over land, low accuracy in presence of dense vegetation or complex topography and difficulties in estimating rainfall in case of soil saturation—the performance of SM2R showed its advantages with the aid of machine learning methods by merging multisource data compared with traditional SM2RAIN datasets.

The relationship between SM fluctuation and rainfall provides an opportunity to retrieve rainfall from the soil moisture. Furthermore, the spatial and temporal variations of the SM pattern also illustrate that the successes of retrieving rainfall largely depends on the underlying characters, including soil types and land cover types. In our next work, we will consider spending more effects to resolve the poor results due to heterogenous underly surface characters; for example, retrieving rainfall from soil moisture in the mix soil and land cover types of pixels.

Future work should either explore other methods and new technologies or expand existing models to better integrate multiple data sources data (i.e., ground/satellite-based SM or rainfall) to produce rainfall products that have improved accuracy. The proposed SM and rainfall product can be used for short-term and long-term forecasting of SM and rainfall or provide reanalysis data for climate and hydrological research and applications such as climate change analysis.

Author Contributions: Conceptualization, Y.Z. and J.H.; methodology, Y.Z.; software, Y.Z.; validation, Y.Z.; formal analysis, Y.Z.; resources, Y.Z. and J.H.; data curation, Y.Z.; writing—original draft preparation, Y.Z.; writing—review and editing, J.H. and C.H.; visualization, Y.Z.; supervision, J.H.; project administration, C.H.; funding acquisition, Y.Z., J.H. and C.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA19040504), the National Natural Science Foundation of China under grants (42130113, 41801271, 41871251, and 41971326), and the Basic Research Innovative Groups of Gansu province, China (Grant No.21JR7RA068).

Data Availability Statement: The SM2R rainfall dataset is available from the authors upon reasonable request.

Acknowledgments: We are grateful for the use of SMOS, SMAP, ASCAT, CMD, CMFD, GPM-SM2RAIN, CHIRPS, ERA5, MSWEP and TMPA data in this study.

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

References

1. Ravuri, S.; Lenc, K.; Willson, M.; Kangin, D.; Lam, R.; Mirowski, P.; Fitzsimons, M.; Athanassiadou, M.; Kashem, S.; Madge, S.; et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature* **2021**, *597*, 672–677. [CrossRef] [PubMed]
2. Kidd, C.; Levizzani, V. Status of satellite precipitation retrievals. *Hydrol. Earth Syst. Sci.* **2011**, *15*, 1109–1116. [CrossRef]
3. Zhang, Y.; Zhang, L.; Hou, J.L.; Gu, J.; Huang, C.L. Development of an Evapotranspiration Data Assimilation Technique for Streamflow Estimates: A Case Study in a Semi-Arid Region. *Sustainability* **2017**, *9*, 1658. [CrossRef]
4. Zhang, Y.; Hou, J.L.; Gu, J.; Huang, C.L.; Li, X. SWAT-Based Hydrological Data Assimilation System (SWAT-HDAS): Description and Case Application to River Basin-Scale Hydrological Predictions. *J. Adv. Model. Earth Syst.* **2017**, *9*, 2863–2882. [CrossRef]
5. Zhang, Y.; Hou, J.H.; Cao, Y.P.; Gu, J.; Huang, C.L. OpenMP parallelization of a gridded SWAT (SWATG). *Comput. Geosci.* **2017**, *109*, 228–237. [CrossRef]
6. Tian, Y.; Peters-Lidard, C.D.; Choudhury, B.J.; Garcia, M. Multitemporal analysis of TRMM-based satellite precipitation products for land data assimilation applications. *J. Hydrometeorol.* **2007**, *8*, 1165–1183. [CrossRef]
7. Huffman, G.J.; Adler, R.F.; Bolvin, D.T.; Gu, G.J.; Nelkin, E.J.; Bowman, K.P.; Hong, Y.; Stocker, E.F.; Wolff, D.B. The TRMM multisatellite precipitation analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *J. Hydrometeorol.* **2007**, *8*, 38–55. [CrossRef]
8. Brocca, L.; Massari, C.; Pellarin, T.; Filippucci, P.; Ciabatta, L.; Camici, S.; Kerr, Y.H.; Fernandez-Prieto, D. River flow prediction in data scarce regions: Soil moisture integrated satellite rainfall products outperform rain gauge observations in West Africa. *Sci. Rep.* **2020**, *10*, 12517. [CrossRef]
9. Hutchinson, M.F. Interpolation of Rainfall Data with Thin Plate Smoothing Splines—Part I: Two Dimensional Smoothing of Data with Short Range Correlation. *J. Geogr. Inf. Decis. Anal.* **1998**, *2*, 153–167.
10. Tarpanelli, A.; Massari, C.; Ciabatta, L.; Filippucci, P.; Amarnath, G.; Brocca, L. Exploiting a constellation of satellite soil moisture sensors for accurate rainfall estimation. *Adv. Water Resour.* **2017**, *108*, 249–255. [CrossRef]
11. Maggioni, V.; Massari, C. on the performance of satellite precipitation products in riverine flood modeling: A review. *J. Hydrol.* **2018**, *558*, 214–224. [CrossRef]
12. Adane, G.B.; Hirpa, B.A.; Lim, C.H.; Lee, W.K. Evaluation and Comparison of Satellite-Derived Estimates of Rainfall in the Diverse Climate and Terrain of Central and Northeastern Ethiopia. *Remote Sens.* **2021**, *13*, 1275. [CrossRef]
13. Fok, H.S.; Chen, Y.T.; Wang, L.; Tenzer, R.; He, Q. Improved Mekong Basin Runoff Estimate and Its Error Characteristics Using Pure Remotely Sensed Data Products. *Remote Sens.* **2021**, *13*, 996. [CrossRef]
14. Rahman, K.U.; Shang, S.H.; Zohaib, M. Assessment of Merged Satellite Precipitation Datasets in Monitoring Meteorological Drought over Pakistan. *Remote Sens.* **2021**, *13*, 1662. [CrossRef]
15. Turk, F.J.; Hristova-Veleva, S.; Giglio, D. Examination of the Daily Cycle Wind Vector Modes of Variability from the Constellation of Microwave Scatterometers and Radiometers. *Remote Sens.* **2021**, *13*, 141. [CrossRef]

16. Ullah, S.; Zuo, Z.K.; Zhang, F.Z.; Zheng, J.H.; Huang, S.F.; Lin, Y.; Iqbal, I.; Sun, Y.Y.; Yang, M.; Yan, L. GPM-Based Multitemporal Weighted Precipitation Analysis Using GPM_IMERGDF Product and ASTER DEM in EDBF Algorithm. *Remote Sens.* **2020**, *12*, 3162. [CrossRef]
17. Sharma, S.; Chen, Y.Y.; Zhou, X.; Yang, K.; Li, X.; Niu, X.L.; Hu, X.; Khadka, N. Evaluation of GPM-Era Satellite Precipitation Products on the Southern Slopes of the Central Himalayas Against Rain Gauge Data. *Remote Sens.* **2020**, *12*, 1836. [CrossRef]
18. Getirana, A.; Kirschbaum, D.; Mandarino, F.; Ottoni, M.; Khan, S.; Arsenault, K. Potential of GPM IMERG Precipitation Estimates to Monitor Natural Disaster Triggers in Urban Areas: The Case of Rio de Janeiro, Brazil. *Remote Sens.* **2020**, *12*, 4095. [CrossRef]
19. Laviola, S.; Monte, G.; Levizzani, V.; Ferraro, R.R.; Beauchamp, J. A New Method for Hail Detection from the GPM Constellation: A Prospect for a Global Hailstorm Climatology. *Remote Sens.* **2020**, *12*, 3553. [CrossRef]
20. Ma, M.H.; Wang, H.X.; Jia, P.F.; Tang, G.Q.; Wang, D.C.; Ma, Z.Q.; Yan, H.M. Application of the GPM-IMERG Products in Flash Flood Warning: A Case Study in Yunnan, China. *Remote Sens.* **2020**, *12*, 1954. [CrossRef]
21. Li, X.; Chen, Y.B.; Deng, X.C.; Zhang, Y.Y.; Chen, L.F. Evaluation and Hydrological Utility of the GPM IMERG Precipitation Products over the Xinfengjiang River Reservoir Basin, China. *Remote Sens.* **2021**, *13*, 866. [CrossRef]
22. Wang, J.X.; Petersen, W.A.; Wolff, D.B. Validation of Satellite-Based Precipitation Products from TRMM to GPM. *Remote Sens.* **2021**, *13*, 1745. [CrossRef]
23. Li, Y.; Wang, W.S.; Wang, G.Q.; Yu, S.Y. Evaluation and Hydrological Application of a Data Fusing Method of Multi-Source Precipitation Products-A Case Study over Tuojiang River Basin. *Remote Sens.* **2021**, *13*, 2630. [CrossRef]
24. North, G.R.; Shen, S.S.P.; Upson, R. Sampling Errors in Rainfall Estimates by Multiple Satellites. *J. Appl. Meteorol.* **1993**, *32*, 399–410. [CrossRef]
25. Zhu, D.T.; Zhang, K.F.; Yang, L.; Wu, S.Q.; Li, L.J. Evaluation and Calibration of MODIS Near-Infrared Precipitable Water Vapor over China Using GNSS Observations and ERA-5 Reanalysis Dataset. *Remote Sens.* **2021**, *13*, 2761. [CrossRef]
26. Dee, D.P.; Uppala, S.M.; Simmons, A.J.; Berrisford, P.; Poli, P.; Kobayashi, S.; Andrae, U.; Balmaseda, M.A.; Balsamo, G.; Bauer, P.; et al. The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.* **2011**, *137*, 553–597. [CrossRef]
27. Wang, L.; Zhang, Y.H.; Yao, Y.J.; Xiao, Z.Q.; Shang, K.; Guo, X.Z.; Yang, J.M.; Xue, S.H.; Wang, J. GBRT-Based Estimation of Terrestrial Latent Heat Flux in the Haihe River Basin from Satellite and Reanalysis Datasets. *Remote Sens.* **2021**, *13*, 1054. [CrossRef]
28. Beck, H.E.; van Dijk, A.; Levizzani, V.; Schellekens, J.; Miralles, D.G.; Martens, B.; de Roo, A. MSWEP: 3-hourly 0.25 degrees global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data. *Hydrol. Earth Syst. Sci.* **2017**, *21*, 589–615. [CrossRef]
29. Jiang, Q.; Li, W.Y.; Fan, Z.D.; He, X.G.; Sun, W.W.; Chen, S.; Wen, J.H.; Gao, J.; Wang, J. Evaluation of the ERA5 reanalysis precipitation dataset over Chinese Mainland. *J. Hydrol.* **2021**, *595*, 125660. [CrossRef]
30. Funk, C.; Peterson, P.; Landsfeld, M.; Pedreros, D.; Verdin, J.; Shukla, S.; Husak, G.; Rowland, J.; Harrison, L.; Hoell, A.; et al. The climate hazards infrared precipitation with stations-a new environmental record for monitoring extremes. *Sci. Data* **2015**, *2*, 150066. [CrossRef]
31. Brocca, L.; Moramarco, T.; Melone, F.; Wagner, W. A new method for rainfall estimation through soil moisture observations. *Geophys. Res. Lett.* **2013**, *40*, 853–858. [CrossRef]
32. Massari, C.; Maggioni, V.; Barbetta, S.; Brocca, L.; Ciabatta, L.; Camici, S.; Moramarco, T.; Coccia, G.; Todini, E. complementing near-real time satellite rainfall products with satellite soil moisture-derived rainfall through a Bayesian Inversion approach. *J. Hydrol.* **2019**, *573*, 341–351. [CrossRef]
33. Wanders, N.; Pan, M.; Wood, E.F. Correction of real-time satellite precipitation with multi-sensor satellite observations of land surface variables. *Remote Sens. Environ.* **2015**, *160*, 206–221. [CrossRef]
34. Ciabatta, L.; Massari, C.; Brocca, L.; Gruber, A.; Reimer, C.; Hahn, S.; Paulik, C.; Dorigo, W.; Kidd, R.; Wagner, W. SM2RAIN-CCI: A new global long-term rainfall data set derived from ESA CCI soil moisture. *Earth Syst. Sci. Data* **2018**, *10*, 267–280. [CrossRef]
35. Kolluru, V.; Kolluru, S.; Wagle, N.; Acharya, T.D. Secondary Precipitation Estimate Merging Using Machine Learning: Development and Evaluation over Krishna River Basin, India. *Remote Sens.* **2020**, *12*, 3013. [CrossRef]
36. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
37. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
38. Teimouri, N.; Dyrmann, M.; Jorgensen, R.N. A Novel Spatio-Temporal FCN-LSTM Network for Recognizing Various Crop Types Using Multi-Temporal Radar Images. *Remote Sens.* **2019**, *11*, 990. [CrossRef]
39. Rajendran, G.B.; Kumarasamy, U.M.; Zarro, C.; Divakarachari, P.B.; Ullo, S.L. Land-Use and Land-Cover Classification Using a Human Group-Based Particle Swarm Optimization Algorithm with an LSTM Classifier on Hybrid Pre-Processing Remote-Sensing Images. *Remote Sens.* **2020**, *12*, 4135. [CrossRef]
40. Gao, W.L.; Gao, J.X.; Yang, L.; Wang, M.J.; Yao, W.H. A Novel Modeling Strategy of Weighted Mean Temperature in China Using RNN and LSTM. *Remote Sens.* **2021**, *13*, 3004. [CrossRef]
41. Zhang, L.; Zhang, J.Z.; Niu, J.; Wu, Q.M.J.; Li, G.S. Track Prediction for HF Radar Vessels Submerged in Strong Clutter Based on MSCNN Fusion with GRU-AM and AR Model. *Remote Sens.* **2021**, *13*, 2164. [CrossRef]
42. Feng, Y.; Zhang, T.S.; Sah, A.P.; Han, L.; Zhang, Z.T. Using Appearance to Predict Pedestrian Trajectories Through Disparity-Guided Attention and Convolutional LSTM. *IEEE Trans. Veh. Technol.* **2021**, *70*, 7480–7494. [CrossRef]

43. Wang, G.Y.; Li, Z.B.; Li, G.J.; Dai, G.Y.; Xiao, Q.; Bai, L.; He, Y.S.; Liu, Y.X.; Bai, S. Real-time liver tracking algorithm based on LSTM and SVR networks for use in surface-guided radiation therapy. *Radiat. Oncol.* **2021**, *16*, 13. [CrossRef] [PubMed]
44. Xie, G.; Niculescu, S. Mapping and Monitoring of Land Cover/Land Use (LCLU) Changes in the Crozon Peninsula (Brittany, France) from 2007 to 2018 by Machine Learning Algorithms (Support Vector Machine, Random Forest, and Convolutional Neural Network) and by Post-classification Comparison (PCC). *Remote Sens.* **2021**, *13*, 3899.
45. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
46. Wang, H.; Ding, Z.D.; Li, X.Y.; Shen, S.Y.; Ye, X.D.; Zhang, D.; Tao, S.F. Convolutional Neural Network with a Learnable Spatial Activation Function for SAR Image Despeckling and Forest Image Analysis. *Remote Sens.* **2021**, *13*, 3444. [CrossRef]
47. Nomura, R.; Oki, K. Downscaling of MODIS NDVI by Using a Convolutional Neural Network-Based Model with Higher Resolution SAR Data. *Remote Sens.* **2021**, *13*, 732. [CrossRef]
48. Hu, L.S.; Qin, M.J.; Zhang, F.; Du, Z.H.; Liu, R.Y. RSCNN: A CNN-Based Method to Enhance Low-Light Remote-Sensing Images. *Remote Sens.* **2021**, *13*, 62. [CrossRef]
49. Liu, J.; Xu, Y.S.; Li, H.H.; Guo, J. Soil Moisture Retrieval in Farmland Areas with Sentinel Multi-Source Data Based on Regression Convolutional Neural Networks. *Sensor* **2021**, *21*, 877. [CrossRef]
50. Srivastava, A.; Saco, P.M.; Rodriguez, J.F.; Kumari, N.; Chun, K.P.; Yetemen, O. The role of landscape morphology on soil moisture variability in semi-arid ecosystems. *Hydrol. Process.* **2021**, *35*, 118C329. [CrossRef]
51. Fatichi, S.; Katul, G.G.; Ivanov, V.Y.; Pappas, C.; Paschalis, A.; Consolo, A.; Kim, J.; Burlando, P. Abiotic and biotic controls of soil moisture spatiotemporal variability and the occurrence of hysteresis. *Water Resour. Res.* **2015**, *51*, 3505–3524. [CrossRef]
52. Wagner, W.; Hahn, S.; Kidd, R.; Melzer, T.; Bartalis, Z.; Hasenauer, S.; Figa-Saldana, J.; de Rosnay, P.; Jann, A.; Schneider, S.; et al. The ASCAT Soil Moisture Product: A Review of its Specifications, Validation Results, and Emerging Applications. *Meteorol. Z.* **2013**, *22*, 5–33. [CrossRef]
53. Kerr, Y.H.; Waldteufel, P.; Wigneron, J.P.; Martinuzzi, J.M.; Font, J.; Berger, M. Soil moisture retrieval from space: The Soil Moisture and Ocean Salinity (SMOS) mission. *IEEE Trans. Geosci. Remote Sens.* **2001**, *39*, 1729–1735. [CrossRef]
54. Chan, S.K.; Bindlish, R.; O'Neill, P.; Jackson, T.; Njoku, E.; Dunbar, S.; Chaubell, J.; Piepmeier, J.; Yueh, S.; Entekhabi, D.; et al. Development and assessment of the SMAP enhanced passive soil moisture product. *Remote Sens. Environ.* **2018**, *204*, 931–941. [CrossRef] [PubMed]
55. 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.; et al. The Soil Moisture Active Passive (SMAP) Mission. *Proc. IEEE* **2010**, *98*, 704–716. [CrossRef]
56. Kun, Y.; Jie, H. *China Meteorological Forcing Dataset (1979–2018)*; National Tibetan Plateau Data Center: Qinghai, China, 2019. [CrossRef]
57. He, J.; Yang, K.; Tang, W.; Lu, H.; Qin, J.; Chen, Y.; Li, X. The first high-resolution meteorological forcing dataset for land process studies over China. *Sci. Data* **2020**, *7*, 25. [CrossRef]
58. Massari, C.; Brocca, L.; Pellarin, T.; Abramowitz, G.; Filippucci, P.; Ciabatta, L.; Maggioni, V.; Kerr, Y.; Fernandez Prieto, D. A daily 25 km short-latency rainfall product for data-scarce regions based on the integration of the Global Precipitation Measurement mission rainfall and multiple-satellite soil moisture products. *Hydrol. Earth Syst. Sci.* **2020**, *24*, 2687–2710. [CrossRef]
59. Li, X.; Cheng, G.D.; Liu, S.M.; Xiao, Q.; Ma, M.G.; Jin, R.; Che, T.; Liu, Q.H.; Wang, W.Z.; Qi, Y.; et al. Heihe Watershed Allied Telemetry Experimental Research (HiWATER): Scientific Objectives and Experimental Design. *Bull. Am. Meteorol. Soc.* **2013**, *94*, 1145–1160. [CrossRef]
60. Vernimmen, R.R.E.; Hooijer, A.; Mamenun; Aldrian, E.; van Dijk, A.I.J.M. Evaluation and bias correction of satellite rainfall data for drought monitoring in Indonesia. *Hydrol. Earth Syst. Sci.* **2012**, *16*, 133–146. [CrossRef]
61. Brocca, L.; Filippucci, P.; Hahn, S.; Ciabatta, L.; Massari, C.; Camici, S.; Schuller, L.; Bojkov, B.; Wagner, W. SM2RAIN-ASCAT (2007–2018): Global daily satellite rainfall data from ASCAT soil moisture observations. *Earth Syst. Sci. Data* **2019**, *11*, 1583–1601. [CrossRef]
62. Gupta, H.V.; Kling, H.; Yilmaz, K.K.; Martinez, G.F. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *J. Hydrol.* **2009**, *377*, 80–91. [CrossRef]
63. Wang, Z.Y.; Li, X.F.; Duan, H.R.; Su, Y.C.; Zhang, X.L.; Guan, X.J. Medical image fusion based on convolutional neural networks and non-subsampled contourlet transform. *Expert Syst. Appl.* **2021**, *171*, 114574. [CrossRef]
64. Stoffelen, A. Toward the true near-surface wind speed: Error modeling and calibration using triple collocation. *J. Geophys. Res. Ocean.* **1998**, *103*, 7755–7766. [CrossRef]
65. Chen, F.; Crow, W.T.; Bindlish, R.; Colliander, A.; Burgin, M.S.; Asanuma, J.; Aida, K. Global-scale evaluation of SMAP, SMOS and ASCAT soil moisture products using triple collocation. *Remote Sens. Environ.* **2018**, *214*, 1–13. [CrossRef]
66. Zhu, G.F.; Li, X.; Zhang, K.; Ding, Z.Y.; Han, T.; Ma, J.Z.; Huang, C.L.; He, J.H.; Ma, T. Multi-model ensemble prediction of terrestrial evapotranspiration across north China using Bayesian model averaging. *Hydrol. Process.* **2016**, *30*, 2861–2879. [CrossRef]
67. Zhang, L.; Ren, D.; Nan, Z.; Wang, W.; Zhao, Y.; Zhao, Y.; Ma, Q.; Wu, X. Interpolated or satellite-based precipitation? Implications for hydrological modeling in a meso-scale mountainous watershed on the Qinghai-Tibet Plateau. *J. Hydrol.* **2020**, *583*, 124629. [CrossRef]
68. Meng, X.J.; Mao, K.B.A.; Meng, F.; Shi, J.C.; Zeng, J.Y.; Shen, X.Y.; Cui, Y.K.; Jiang, L.M.; Guo, Z.H. A fine-resolution soil moisture dataset for China in 2002–2018. *Earth Syst. Sci. Data* **2021**, *13*, 3239–3261. [CrossRef]

69. Chen, W.; Shen, H.; Huang, C.; Li, X. Improving Soil Moisture Estimation with a Dual Ensemble Kalman Smoother by Jointly Assimilating AMSR-E Brightness Temperature and MODIS LST. *Remote Sens.* **2017**, *9*, 273. [CrossRef]
70. Huang, C.; Chen, W.; Li, Y.; Shen, H.; Li, X. Assimilating multi-source data into land surface model to simultaneously improve estimations of soil moisture, soil temperature, and surface turbulent fluxes in irrigated fields. *Agric. Meteorol.* **2016**, *230–231*, 142–156. [CrossRef]
71. Ma, L.; Liu, Y.; Zhang, X.L.; Ye, Y.X.; Yin, G.F.; Johnson, B.A. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS J. Photogramm.* **2019**, *152*, 166–177. [CrossRef]
72. Zhang, W.J.; Xue, X.; Peng, F.; You, Q.G.; Hao, A.H. Meta-analysis of the effects of grassland degradation on plant and soil properties in the alpine meadows of the Qinghai-Tibetan Plateau. *Glob. Ecol. Conserv.* **2019**, *20*, e00774. [CrossRef]