remote sensing

MDPI

Article

A Machine Learning-Based Approach for Surface Soil Moisture Estimations with Google Earth Engine

Felix Greifeneder [1,*], Claudia Notarnicola [1] and Wolfgang Wagner [2]

[1] Institute for Earth Observation, Eurac Research, 39100 Bolzano, Italy; claudia.notarnicola@eurac.edu
[2] Department of Geodesy and Geoinformation, TU Wien, 1040 Vienna, Austria; wolfgang.wagner@geo.tuwien.ac.at
* Correspondence: felix.greifeneder@eurac.edu

Abstract: Due to its relation to the Earth's climate and weather and phenomena like drought, flooding, or landslides, knowledge of the soil moisture content is valuable to many scientific and professional users. Remote-sensing offers the unique possibility for continuous measurements of this variable. Especially for agriculture, there is a strong demand for high spatial resolution mapping. However, operationally available soil moisture products exist with medium to coarse spatial resolution only ($\geq$1 km). This study introduces a machine learning (ML)—based approach for the high spatial resolution (50 m) mapping of soil moisture based on the integration of Landsat-8 optical and thermal images, Copernicus Sentinel-1 C-Band SAR images, and modelled data, executable in the Google Earth Engine. The novelty of this approach lies in applying an entirely data-driven ML concept for global estimation of the surface soil moisture content. Globally distributed in situ data from the International Soil Moisture Network acted as an input for model training. Based on the independent validation dataset, the resulting overall estimation accuracy, in terms of Root-Mean-Squared-Error and R^2, was 0.04 $m^3 \cdot m^{-3}$ and 0.81, respectively. Beyond the retrieval model itself, this article introduces a framework for collecting training data and a stand-alone Python package for soil moisture mapping. The Google Earth Engine Python API facilitates the execution of data collection and retrieval which is entirely cloud-based. For soil moisture retrieval, it eliminates the requirement to download or preprocess any input datasets.

Keywords: soil moisture; Sentinel-1 SAR; Landsat-8 optical/thermal data; machine learning; cloud-based approach; Google Earth Engine

Citation: Greifeneder, F.; Notarnicola, C.; Wagner, W. A Machine Learning-Based Approach for Surface Soil Moisture Estimations with Google Earth Engine. *Remote Sens.* **2021**, *13*, 2099. https://doi.org/10.3390/rs13112099

Academic Editor: Lefei Zhang

Received: 7 May 2021
Accepted: 22 May 2021
Published: 27 May 2021

1. Introduction

The soil moisture content (SMC) is a crucial state variable in the complex global cycles of water, energy, and carbon, and is therefore very relevant for studying the Earth's climate and weather [1]. Furthermore, SMC plays a crucial role in natural hazards like drought, floods, and landslides [2]. Satellite remote sensing presents the only possibility for the spatially continuous measurement of surface SMC over large areas. Current, widely used approaches belong to two main categories: those based on active or passive microwave remote sensing, and those based on optical (i.e., shortwave and thermal radiation) remote sensing. The underlying methods for the estimation of SMC are fundamentally different. Most microwave-based retrieval algorithms rely on the same principle, exploiting the dielectric properties of water and its effect on the reflected microwave radiation [3]. For optical remote sensing, many different approaches exist, exploiting the relationship between SMC and surface reflectance, changes of vegetation indices, or surface temperature [4]. Independent of the type of underlying measurement, these approaches require complex retrieval models, often relying on assumptions and approximations.

The essential advantages of microwaves are their low sensitivity to atmospheric conditions, sun-illumination and clouds, and the fact that there is a direct, physical relationship

between the moisture content of soil and the emitted and reflected energy [3]. However, measurements are also strongly influenced by vegetation water content and structure and surface roughness, which can be challenging to estimate. There are two main groups of traditional modelling approaches: process-based models, encompassing physical models like the integral equation model (IEM) [5]; semi-empirical models, e.g., from Oh et al. [6] and Dubois et al. [7]; and change detection approaches, e.g., from Wagner et al. [8]. For optical-based SMC estimations, many different approaches, distinguished by the used frequency, exist. The main advantages of these methods are many existing optical satellites and good data availability at different spatial and spectral resolutions. Methods based on visible light and near-infrared often exploit the surface reflectance properties of bare soil, for which studies have shown a negative correlation of the reflectance in water absorption bands and SMC [9]. Due to this dependency, various drought indices, like the normalized-difference-vegetation-index (NDVI) or the normalized-difference-water-index (NDWI), can serve as a proxy. The estimation of SMC based on thermal remote sensing has been the subject of a large number of studies over the last few decades, which led to the development of several different models and approaches, exploiting the relationship to land-surface-temperature (LST) SMC [10–12]. The third category of approaches uses the combination of LST and vegetation indices. Price [13] introduced the often applied, so-called triangle model, which uses the distribution of values in a triangular representation of the LST-vegetation index feature space to obtain SMC. The main drawback of optical remote sensing approaches, compared to microwave-based approaches, is their sensitivity to atmospheric conditions and clouds.

Currently available operational SMC products rely on data from coarse- to medium-resolution passive or active microwave sensors like SMOS [14] (passive), SMAP [15] (passive), or ASCAT [16] (active). As a component of the Copernicus Land Monitoring Service, the Soil Water Index [17], based on a fusion of Sentinel-1 (S1) and ASCAT data, offers medium resolution (1 km) SMC observations. The advantage of these coarse- to medium-resolution sensors is their high temporal resolution, offering daily observations. With S1, Sentinel-2 and Landsat-8 high resolution (<50 m) remote-sensing data, in the microwave as well as the shortwave and thermal domains, are available operationally and on an open access basis, providing the most important foundation for high resolution SMC mapping [18,19].

Beyond the more traditional modelling approaches described above, machine learning (ML) offers some alternative approaches. Due to the high complexity of physical models, the popularity of ML for the remote-sensing based estimation of biophysical parameters has grown significantly over the last decade [20]. The flexibility of ML approaches is further highlighted by their potential to be used in various hydro-meteorological applications, from the prediction of SMC to precipitation, temperature, or wind [21–25]. Compared to more traditional approaches, these methods have two significant advantages: (1) they enable the construction of more objective, purely data-driven retrieval models, independent of necessary assumptions; and (2) they allow the combination of data from different sources (like the combination of optical and microwave remote sensing), exploiting their relationship with a target variable.

There are various ways to incorporate ML in the estimation approach. Often, it is applied for model inversion or downscaling purposes. Moosavi et al. [26] presented an example of how these concepts can be applied to MODIS and Landsat imagery for the downscaling of land-surface-temperature measurements and the estimation of high-resolution (100 m) SMC by applying support-vector-regression (SVR) and an adaptive neuro-fuzzy inference system for model inversion. Srivastava et al. [27] applied an artificial neural network (ANN) to downscale a SMOS product based on MODIS imagery. In the context of microwave remote sensing, ML has been used to exploit high-resolution Synthetic Aperture Radar (SAR) data for high-resolution mapping purposes [28–32]. In most cases, these studies focused on a specific region or study area. Kolassa et al. [33] were able to demonstrate the effectiveness of a data-driven approach, for building a more

generally applicable model, by modelling SMC with an ANN using coarse resolution (36 km) SMAP data as an input. ML presents an effective way to combine or fuse different types of data (e.g., from remote sensing, in situ measurements, or models). Results from a study by Liu et al. [34] show that the SMC estimation accuracies, for a farmland test area, by a combination of Sentinel-1 and Sentinel-2 and several different ML algorithms (SVR, deep neural networks, and generalized regression neural networks) were higher than those by traditional semi-empirical models based on either Sentinel-1 or Sentinel-2. Another example for data fusion was presented by Bhuyian et al. [35], with a study in which SMAP data were combined with MODIS data in an ML model to improve the estimation of precipitation.

The increasing availability of open data (e.g., from the Copernicus program), paired with the emergence of platforms like Google Earth Engine (GEE), which offer analysis-ready data and server-side processing capabilities, has further increased the popularity of ML for the estimation of biophysical parameters in recent years [36–38]. Due to the high volume of data from modern satellite missions, access and processing have become more complex [39], which means that this shift in the data exploitation strategies has become necessary.

The study presented hereafter followed a similar approach to that of Pasolli et al. [30], who used in situ SMC measurements as a target variable for the ML algorithm and the construction of an empirical model. One of the novel aspects of this work is to propose a data-driven approach that can be applied regardless of location, to be globally applicable but locally relevant. Instead of focusing on a specific study region like in previous studies, the model was trained and tested on the International Soil Moisture Network (ISMN), with global coverage. Chatterjee et al. followed a similar aim with the work presented in [38]. They introduced an approach to training a model for SMC estimations within the entire continental United States of America using US Climate Reference Network (USCRN) measurements as a training target. Some limitations described in [40] were related to the lack of accurate auxiliary data (e.g., land cover, soil type). The article presented here demonstrates how some of these limitations can be overcome by (1) further increasing the size of the training dataset using measurements from the International Soil Moisture Network (ISMN) and (2) combining Sentinel-1 SAR data with optical data from Landsat-8. It shows how ML can be applied in a data-driven approach to estimate high-resolution SMC based on a spatially dispersed training dataset. The proposed solution tackles a gap of the currently available operational datasets regarding their spatial resolutions. Existing operational satellite-based soil moisture products focus on the mapping at medium to coarse spatial resolutions [14–17]. Furthermore, one of the study outputs is a software toolbox allowing the fully cloud-based mapping of the SMC, enabling easy access for data users and integration in other studies.

2. Data and Study Area

The following section describes the datasets used in the present study, i.e., in situ data of the ISMN Network [41,42], S1 backscatter measurements, Landsat-8 (L8) shortwave reflectance and thermal radiance, and modelled surface parameters from the global land-surface model GLDAS. The analysis focuses on the period from October 2014 until mid-2020. Google Earth Engine (GEE) provided all datasets except ISMN. The training set encompassed approximately 30,000 samples.

The study area extent is, in principle, global (restricted to specific land-cover classes). Based on the Copernicus Global Land Layer (see Section 2.2.2), masking was carried out to include only the following classes (Section 3.2 describes the masking process in more detail): shrubs, herbaceous vegetation, cropland, bare/sparse vegetation, open forest.

2.1. The International Soil Moisture Network

The ISMN is an initiative to establish and maintain a global in situ soil moisture database. Through its website (https://ismn.geo.tuwien.ac.at, accessed on 28 February

2020), data from hundreds of monitoring stations worldwide from many providers are available. This network's primary goal is to provide the basis for the large-scale validation of satellite-derived soil moisture products. All data provided through the ISMN are free to use for scientific purposes.

The ISMN dataset is heterogenous, i.e., stations are operated by different providers, using various measurement techniques at different depths, and the stations are located in different land-cover types and climate zones. Therefore, some of the ISMN data may not be suitable because it represents SMC variability that cannot be detected by satellite remote sensing. Due to the large number of stations, individual selection was impossible. Therefore, based on the known limitations of used satellite datasets, a set of rules was defined to filter the dataset (Section 3.2 describes the masking procedure). The number of 461 globally distributed ISMN stations from 13 in situ networks met the requirements to provide target measurements for the algorithm training. Most of the stations are located in North America. Table 1 shows a list of the used monitoring networks (with their location, provider, and available references) and Figure 1 their geographic distribution. The remaining measurement errors and uncertainties are part of the inherent noise, and contribute to the general uncertainty of the retrieval model. The open source python package pytesmo [43] provides reading and postprocessing tools for the ISMN dataset.

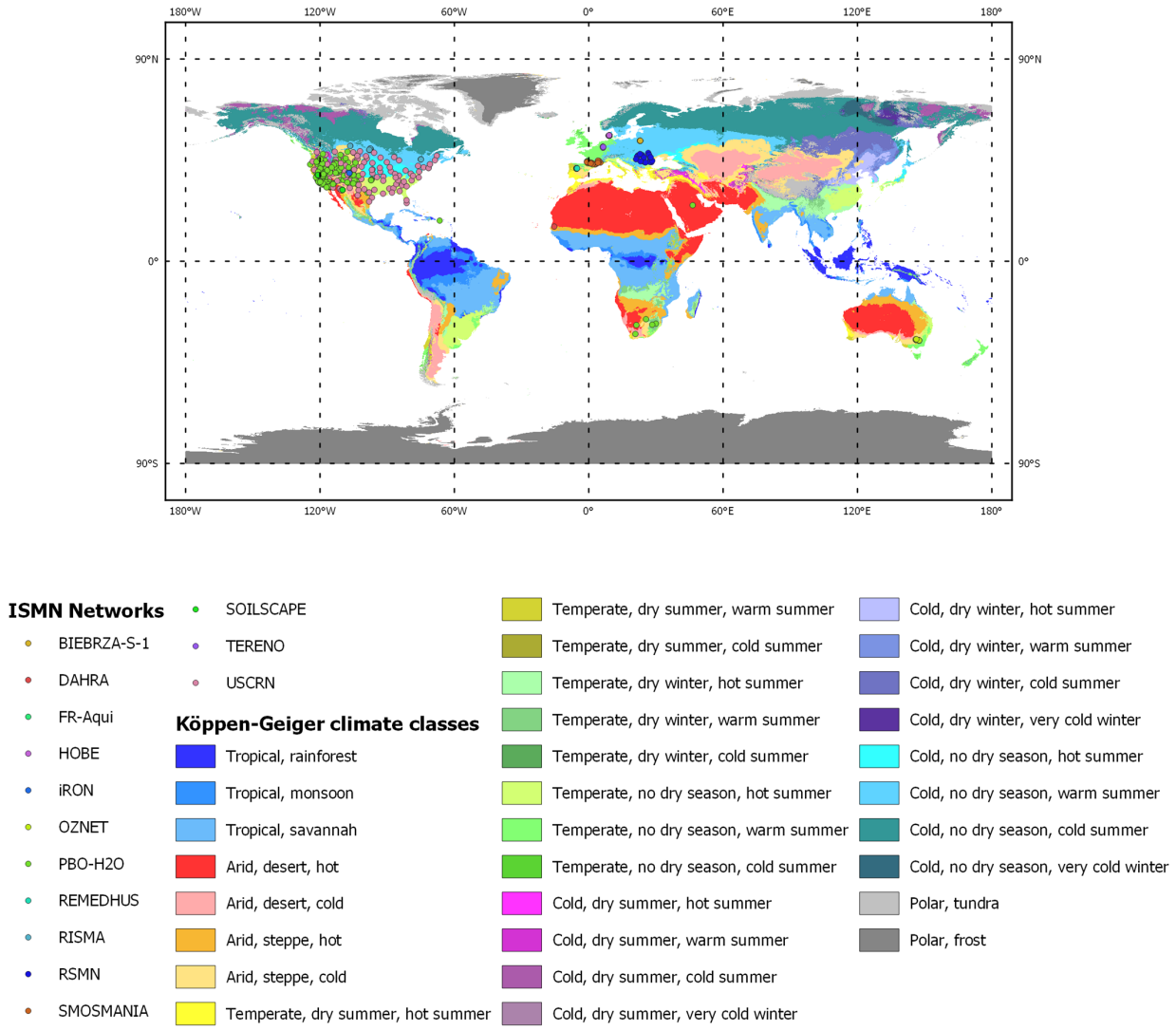

Figure 1. A map with the locations of the ISMN soil moisture monitoring sites that the study uses. The Köppen-Geiger climate classification [44] background map shows the global distribution of climate zones.

Table 1. A list of the soil moisture networks of which provide data to train the SMC estimation model.

Network Name	No. of Stations Used in the Study	Country	Provider	Reference
USCRN	102	USA	NOAA NCDC	[45]
OZNET	12	Australia	University of Melbourne	[46,47]
PBO-H2O	138	USA	University of Colorado	[48]
SOILSCAPE	80	USA	University of Southern California	[49,50]
HOBE	25	Denmark	Hydrological Observatory	[51]
SMOSMANIA	20	France	CNRM/GAME, METEO-FRANCE, CRNS	[52,53]
REMEDHUS	20	Spain	Universidad de Salamanca	-
RSMN	18	Romania	National Meteorological Administration	-
FR-Aqui	2	France	Institute of Agricultural Research	-
BIEBRZA-S-1	18	Poland	Instytut Geodezji i Kartografii	-
RISMA	13	Canada	Agriculture and Agri-Food Canada	[54]
iRON	8	USA	Aspen Global Change Institute	[55]
TERENO	4	Germany	Helmholtz Gemeinschaft Forschungszentrum Jülich	[56]
DAHRA	1	Senegal	Copenhagen University	[57]

2.2. Google Earth Engine

S1 A and B produce more than 1 TB of data daily [58]. During its operational lifespan, S1 will generate an enormous amount of data. Consequently, the user would have to handle a substantial volume of data and invest a significant amount of time for the preprocessing of low-level satellite data to fully exploit the potential of the S1 archive. For many other satellite, model-based, or geospatial datasets, users are facing the same problems. This problem has sparked a paradigm shift for the large-scale analytics of geospatial datasets. Recent years have shown the emergence of more and more providers offering cloud processing and online access to analysis-ready data. One of these platforms is GEE. GEE hosts the satellite and auxiliary data for this study. The GEE Python Application Programming Interface (API) allows convenient access to its data and processing functionality [39].

2.2.1. Sentinel-1

S1 is a C-Band Synthetic Aperture Radar (SAR) operated within the Copernicus program, which is a joint initiative of the European Commission (EC) and the European Space Agency (ESA). The standard acquisition mode over land is the Interferometric Wide Swath Mode (IW), with acquisitions at a 250 km wide swath and a spatial resolution of 5 by 20 m. S1 flies in a near-polar, sun-synchronous orbit with a 12-day repeat cycle. The two satellites A and B share the same orbit plane with a 180° orbital phasing difference, which results in a 6-day repeat cycle for the S1 constellation. A description of all sensor and platform details can be found in ESAs S1 user handbook [59]. The data available on GEE provide σ^0 based on the dual-polarization (VV + VH) Ground Range Detected (GRD) product.

2.2.2. Copernicus Global Land Cover Layer (CGLS-LC100)

The CGLS-LC100 [60] delivers global-land cover maps at a spatial resolution of 100 m, derived from optical satellite remote sensing data. These include a discrete classification and the fractional cover for specific land-cover types. These maps are updated annually, starting from 2015. Land-cover data provided input for masking and were a feature candidate for the SMC retrieval model.

2.2.3. The Global Land Data Assimilation System (GLDAS)

GLDAS [61] ingests satellite and ground-based observational data products and uses advanced land surface modelling and data assimilation techniques to simulate many land-surface parameters. The dataset in version 2.1 covers the period from 1 January 2000, to the present, with about one month latency. Soil temperature and snow-water-equivalent

(SOILTMP0_10cm_inst and SWE_inst) were required to mask in situ measurements and satellite data.

2.2.4. Landsat-8 Shortwave Reflectance and Thermal Radiance

L8 is a satellite operated by the USGS, providing imagery of the entire Earth every 16 days with a spatial resolution of 30 m to 100 m. It is acquiring data using two instruments, the Operational Land Imager (OLI) and the Thermal Infrared Sensor (TIRS). Data in the GEE data collection USGS L8 Surface Reflectance Tier 1, which consists of atmospherically corrected surface reflectance for five visible and near-infrared (NIR) bands, two short wave infrared bands, and two thermal infrared bands [62], were considered as potential features for the retrieval model (Table 2).

Table 2. Overview of the Landsat-8 bands with their respective spatial resolution and spectral width.

Landsat-8 Bands	Spatial Resolution [m]	Spectral Width [μm]
Band 1—Coastal/Aerosol	30	0.435–0.451
Band 2—Blue	30	0.452–0.512
Band 3—Green	30	0.533–0.590
Band 4—Red	30	0.636–0.673
Band 5—Near-Infra-Red	30	0.851–0.879
Band 6—Shortwave-Infrared-1	30	1.566–1.651
Band 7—Shortwave-Infrared-2	30	2.107–2.294
Band 10—Thermal-Infrared-1	100	10.60–11.19
Band 11—Thermal-Infrared-2	100	11.50–12.51

2.2.5. MOD13Q1 Enhanced Vegetation Index

The MODIS Enhanced Vegetation Index (EVI) is an improved version of the Normalized Difference Vegetation Index (NDVI), which minimizes canopy background variations and maintains better sensitivity over dense vegetation conditions by including also the blue band and information about atmospheric influences. The MOD13Q1.005 [63] product provides 16-day temporal composites at a spatial resolution of 250 m with global coverage. With its consistent temporal information, the EVI complemented L8 to capture vegetation dynamics.

2.2.6. OpenLandMap (OLM) Soil Information

The spatial distribution and patterns of SMC have a strong link to soil properties. The soil texture class [64], soil bulk density [65], clay content [66], and sand content [67] for the 0 cm layer of the OLM collection were acting as training feature candidates. This dataset provides maps with global coverage at 250 m spatial resolution.

3. Methods

This section describes the applied methods, covering feature extraction, data masking and merging, data preprocessing, and the ML model training (Figure 2). The subsections describe the individual steps in more detail.

3.1. S1 Preprocessing

Several SAR preprocessing steps were already pre-applied to the data, which are available on GEE [68]:

1. Apply orbit file
2. Thermal noise removal
3. Radiometric calibration
4. Terrain correction using the SRTM 30, or the ASTER DEM for areas beyond ±60° latitude, where SRTM is not available
5. Resampling to a 10 m grid

Beyond these steps above, the following custom processing steps were applied:

1. Multitemporal speckle filter
2. Radiometric terrain correction
3. Feature extraction

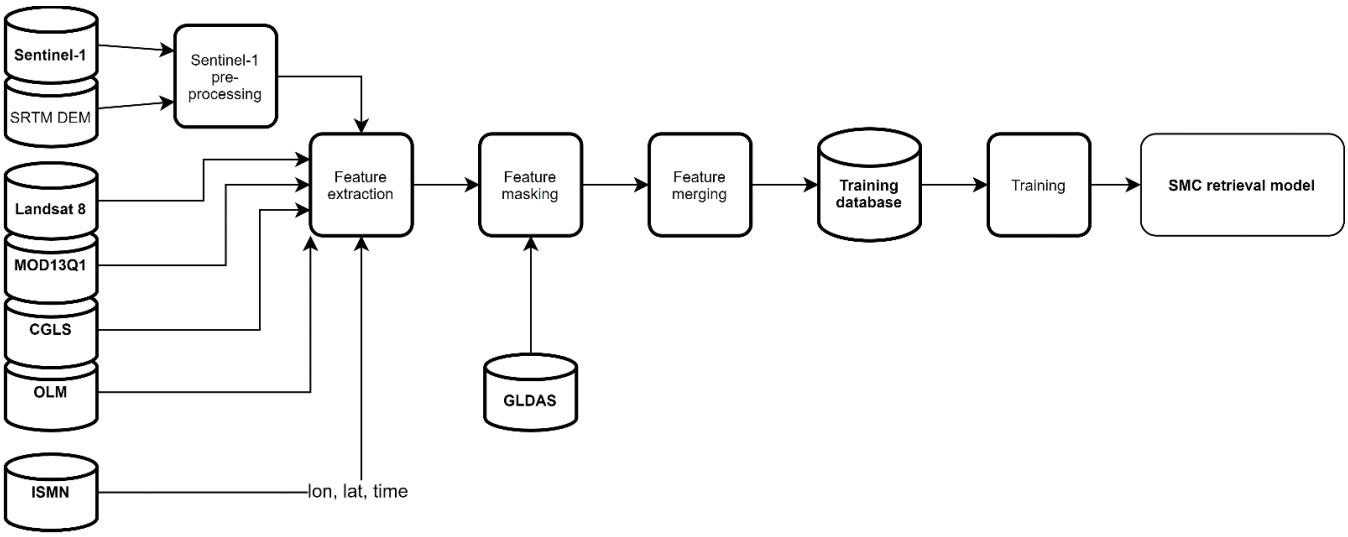

Figure 2. Flowchart with an overview of the processing steps for feature extraction, masking, and merging.

3.1.1. Multitemporal Speckle Filter

SAR imagery is prone to a specific type of noise, i.e., the characteristic salt and pepper effect called speckle. Speckle is not a type of noise introduced by measurement system errors, but by the actual physical properties of the observed target, which cause a noise-like signal. It results from constructive and deconstructive interference of the coherent, but phase-shifted (caused by the random orientation of subpixel scatterers) returned signal. The GEE preprocessing does not include any speckle filtering. Quegan et al. [69] demonstrated the effectiveness of a multitemporal approach that incorporates both the spatial and temporal pixel neighborhood to determine the local value. The corrected intensity J at the pixel location (x, y) for image k, with the original pixel value I, and the spatial average of the pixel neighborhood denoted as <.> is defined as

$$J_k(x,y) \;=\; \frac{I_k}{N} \sum_{i=1}^{N} \frac{I_i(x,y)}{I_i},\; 1 \le k \le N. \tag{1}$$

3.1.2. Radiometric Terrain Correction

Due to the side-looking viewing geometry of S1, the data are particularly affected by topography, causing geometric and radiometric distortions. Vollrath et al. [70] developed a GEE based approach for the angular-based radiometric slope correction to correct these distortions. It employs two reference models, for volume- or surface-scattering dominated surfaces, computing the radiometrically corrected γ^0_f. The following equation is applied, in case of volume-scattering [71]:

$$\gamma^0_f \;=\; \gamma^0 \frac{\tan(90-\theta_i)}{\tan(90-\theta_i+\alpha_r)}, \tag{2}$$

with the slope steepness in range direction α_r and the incidence angle θ_i. In case of surface-scattering case, the approach exploits a model by Ulander [72],

$$\gamma^0_f \;=\; \gamma^0 \frac{\cos(\alpha_{az})\cos(90-\theta_i+\alpha_r)}{\cos(90-\theta_i)}, \tag{3}$$

which adds the tilt in azimuth direction α_{az} as an additional quantity. The incidence angle corrected backscatter γ^0 is derived from the normalized radar cross-section σ^0:

$$\gamma^0 = \frac{\sigma^0}{\cos\theta_i}. \quad (4)$$

The Shuttle Radar Topography Mission [73] (SRTM) digital elevation model (DEM) V3 provides the topographic reference.

3.1.3. Computation of Temporal Statistics

For bare soil, the backscatter intensity is determined mainly by SMC and surface roughness [3]. The structure and water content of vegetation, if present, create a signal, which adds to that of the soil. In [74], the authors demonstrated that in extreme cases, caused, for example, by the row patterns in agricultural fields, the effect of roughness could be as strong as 10 dB, dominating the SMC signal by far. The effect depends strongly on the local incidence angle but on average, a roughness related σ^0 variability of not less than 2 dB can be expected. Furthermore, in an experiment based on C-Band VV data from ERS-1, the sensitivity of σ^0 to SMC was quantified as 0.26 dB/0.01 $m^3 \cdot m^{-3}$, which shows that roughness information is crucial for the retrieval of SMC. For the change detection approach, Wagner et al. [75] assume that surface roughness remains constant over time or that changes occur on very large time scales, i.e., it is responsible for a constant background signal. Temporal backscatter statistics, median and second, third, and fourth central statistical moments were computed to characterize these static effects caused by surface roughness.

3.2. Feature Extraction, Masking, and Merging

The feature extraction was carried out in GEE based on the ISMN station locations and the measurements dates. Each dataset was resampled to a spatial resolution of 50 m using bilinear interpolation. Values were then extracted based on the average of all pixels touched by a 50 m diameter circular buffer around the sampling location. After applying the filtering criteria listed in Table 3, the training database contained approximately 30,000 samples in 62 features (Table 4)

Table 3. Summary of the filtering criteria applied to the training database.

Variable	Valid If
ISMN temporal overlap	>0
ISMN sensing depth	≤5 cm
S1 layover and foreshortening masks [56]	0
L8 pixel quality band	Not affected by clouds, terrain occlusion, or radiometric saturation
EVI	<0.5
CGLS-LC100 class	20, 30, 40, 60, 121, 122, 124, 125, 126 [1]
GLDAS SoilTMP0_10cm_inst	>275 K
GLDAS SWE_inst	0 kg/m^2

[1] 20: shrubs; 30: herbaceous vegetation; 40: cropland; 60: bare/sparse vegetation; 121–126: open forest.

Table 4. The complete list of features, which were extracted from the input datasets.

#	Feature Name	#	Feature Name
1	S1 $\gamma^0_{VV}vol.$	22	% of moss and lichen
2	S1$\gamma^0_{VH}vol.$	23	% of urban areas
3	S1$\gamma^0_{VV}surf.$	24	% of permanent water bodies
4	S1$\gamma^0_{VH}surf.$	25	% of seasonal water bodies
5–13	Median and standard deviation S1 $\gamma^0_{VV}vol.$, S1 $\gamma^0_{VH}vol.$, S1 $\gamma^0_{VV}surf.$, S1 $\gamma^0_{VH}surf.$	26–35	L8, bands 1–7; 10–11

Table 4. *Cont.*

#	Feature Name	#	Feature Name
14	S1 Local-Incidence-Angle	36	No. of days between S1 and L8 acquisitions
15	S1 orbit direction (ascending/descending)	37–55	Median and standard deviation L8, bands 1–7; 10–11
16	Land-cover class	56	MODIS EVI
17	% of bare areas	57–58	Median and standard deviation MODIS EVI
18	% of crop areas	59	OLM bulk density
19	% of tree cover	60	OLM clay content
20	Forest type	61	OLM sand content
21	% of grassland	62	OLM texture class

3.3. Model Training

Studies [76,77] show that the choice of the best ML algorithm for retrieving biophysical parameters varies significantly, depending on the target variable and the structure of the training dataset. This study used a Gradient Boosted Regression Trees (GBRT) algorithm. GBRT belongs to the ensemble methods, which means several weak learners are built and combined into one powerful ensemble. Weak learners are combined sequentially, and each newly added model tries to correct the prediction bias of all previous models combined [78,79]. GBRT, like Random Forest, belongs to the family of tree-based methods, which means that it is naturally compatible with different data types and ordinal scales. Further advantages are its insensitivity to differently scaled features and the low computational cost associated with algorithm training and target prediction. Moreover, it has proven to perform very well in similar applications [80–82]. A comparison of GBRT with three other popular ML methods is provided in supplementary document S1. GBRT uses an additive model to combine weak learners:

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x), \quad (5)$$

where M is the total number of models, γ is a multiplication factor (the so-called step size), and $h(x)$ is the weak learner. An iterative process is building the model using decision trees of fixed size as weak learners. The loss L is minimized (with a least-square loss function, in this case) for each tree h_m given the previous ensemble F_{m-1}:

$$h_m = \underset{h}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + h(x_i)), \quad (6)$$

where x_i and y_i are the feature and target values of the training dataset. GBRT solves the minimization problem numerically via steepest descent [83]. The following equation determines the value of γ, based on line-search [84]:

$$\gamma_m = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L\left(y_i, F_{m-1}\left(x_i - \gamma \frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)}\right)\right). \quad (7)$$

An in-depth discussion of GBRT is beyond the scope of this article. For further details, the reader should refer to [85]. This study used the algorithm implementation of Scikit-Learn [86]. Figure 3 shows the detailed workflow of the training procedure, which consists of the following steps:

1. Generation of training and test datasets: the test dataset consisted of 20% of the samples and was randomly selected. The remaining 80% represented the training dataset.
2. Tuning and training: cross-validation (CV) and a grid-search approach drove the optimization of hyperparameters [87]. To be specific, Leaf-One-Group-Out-Cross-Validation (LOGO-CV) was applied to find the optimal setting of hyperparameters. Table 5 shows which parameters were tuned and how the search space was defined. The hyperparameters, which are not listed in the table, were set to the default values of the Scikit-Learn implementation. Following the LOGO approach for calculating the validation score, the training dataset was split up into N groups (according to

the 461 ISMN stations). Iteratively, the algorithm training was performed based on data from N-1 groups and validated against the left-out group. The average based on all iterations gave the final score. The validation score was calculated for every possible hyperparameter configuration. The LOGO-CV approach allows for estimating the trained model's generalization capabilities, favoring hyperparameter settings, configuring the algorithm to be less prone to overfitting. The grouping of samples based on the ISMN stations simulates the estimation model's application for an unseen location.

3. Feature selection: the training procedure was further nested in an automatic feature selection routine wherein the training was repeated iteratively with the least essential feature removed in each step. The ranking was based on the impurity-based feature importance, which is provided as an output of the Scikit-Learn implementations.
4. Testing: the final assessment of the SMC estimation accuracy (the test score) was performed based on the independent test dataset extracted from the whole dataset before the training procedure. Therefore, the test dataset constituted an unseen set and allowed estimating the model's generalization capabilities. A sizeable negative difference between the validation score and the test score would indicate overfitting of the model to the training dataset.

Table 5. Definition of the search space for hyperparameter tuning.

Algorithm	Hyperparameters	
GBRT	Learning-rate	0.01, 0.1, 0.2
	Number of estimators	100, 500, 1000
	The fraction of samples used for fitting	0.2, 0.5, 1
	Maximum depths of individual regression trees	3, 5, 10
	Early stopping after *n* iterations with no change	10

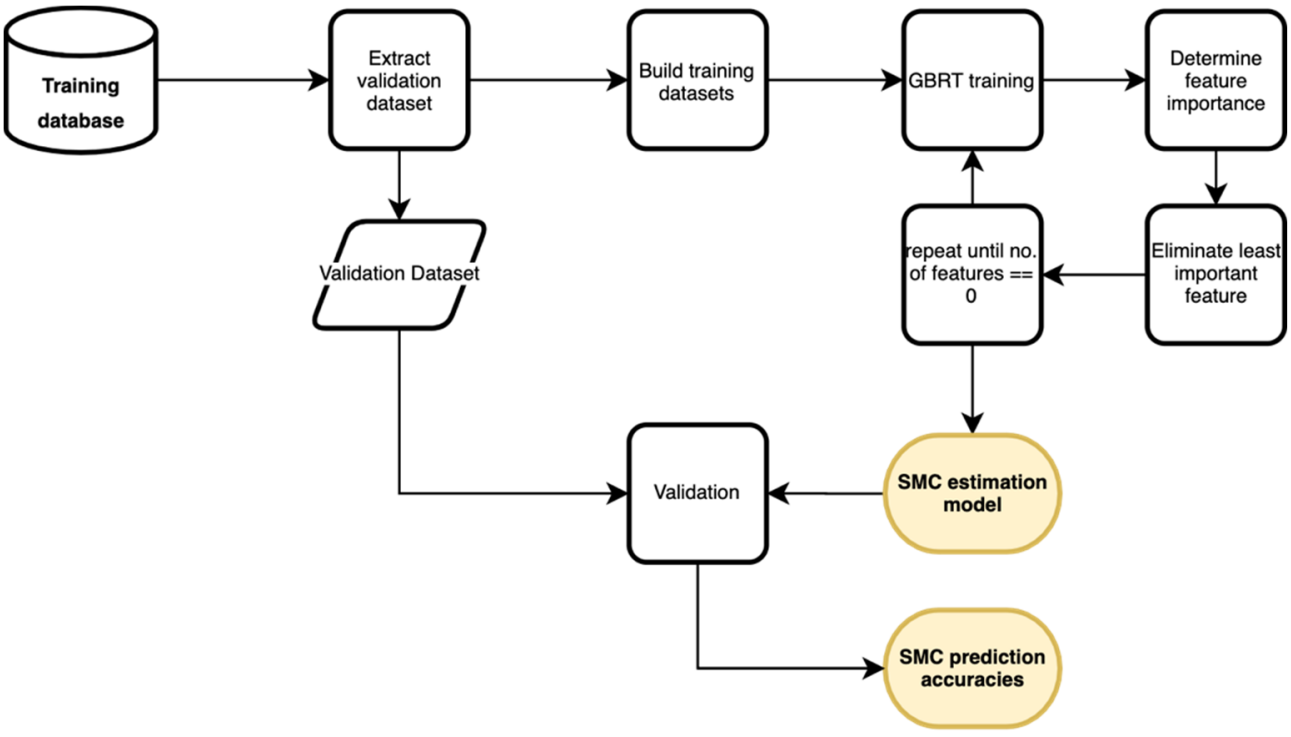

Figure 3. This flowchart describes the steps, which were carried out during the model training procedure.

3.4. Implementation

For the training, ISMN data were prefiltered, downloaded and stored locally. Based on the in situ locations (in space and time), the S1, L8, GLDAS and other auxiliary data were retrieved from GEE using the Python API. All preprocessing and filtering steps were carried out server-side, i.e., only the filtered and preprocessed training data were downloaded to the local workstation. Scikit-learn [86] was used to solve the regression problem and derive the GBRT model. Figure 4 gives a schematic overview of the architecture. Depending on the number of training samples, the training procedure can be time-consuming, mainly due to GEE's data retrieval. The GEE based mapping of SMC was implemented as a stand-alone Python package, the PYthon Sentinel-1 Soil-Moisture Mapping Toolbox (PYSMM) [88]. PYSMM is freely available as open-source software and was delivered as Supplementary Material to this article.

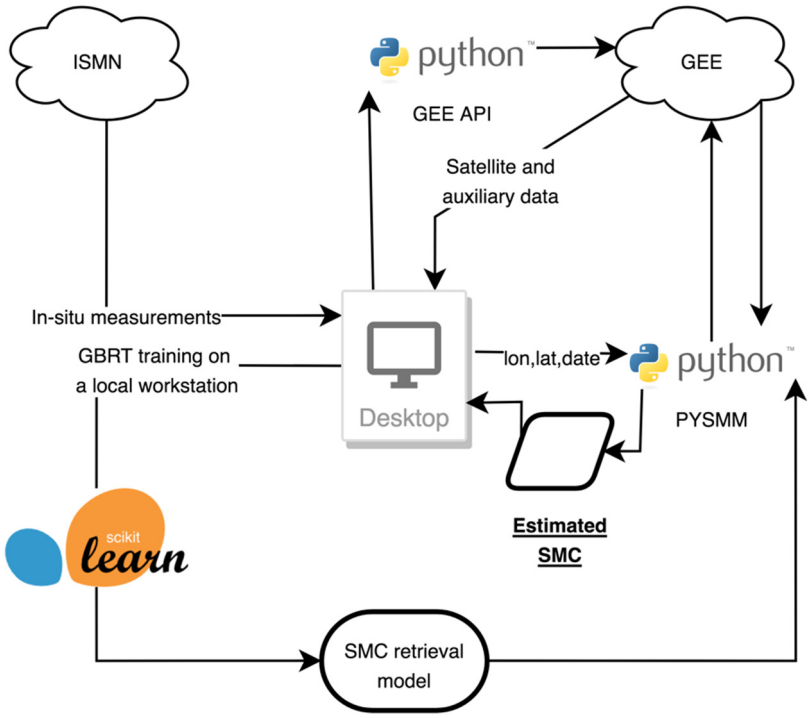

Figure 4. The implementation of the retrieval and estimation procedure: data extraction from GEE, offline model training, and the online estimation module.

4. Results and Discussion

The following chapters present the algorithm training results, assess the GBRT performance, and evaluate the SMC validation results.

4.1. Algorithm Training and Validation Results

Table 6 presents the results of the LOGO-CV based hyperparameter selection described in Section 3.3.

Table 6. Result of the hyperparameter selection.

Algorithm	Hyperparameters	
GBRT	Learning-rate	0.1
	Number of estimators	100
	The fraction of samples used for fitting	0.5
	Maximum depths of individual regression trees	10
	Early stopping after *n* iterations with no change	10

A summary of validation and test scores is presented in Table 7. The similarity of the coefficient of determination (R^2) and the Root-Mean-Squared-Error (RMSE) based on the cross-validation average and based on the test set dataset shows that the GBRT model is not subject to overfitting. It is noteworthy that a model with a relatively low level of complexity (100 base estimators and a maximum depth of the individual trees of 10) achieved the best results. Fewer and shallower regression trees lead to a low computational cost of model training and target estimation.

Table 7. The validation and test-score for the estimation model.

Method	LOGO R^2	LOGO RMSE [$m^3 \cdot m^{-3}$]	Test-Set R^2	Tet-Set RMSE [$m^3 \cdot m^{-3}$]	Training Time * [s]	Prediction Time (Test-Set) [s]
GBRT	0.726	0.054	0.812	0.044	5.20	0.05

* Excluding LOGO-CV and feature selection.

The sound overall predictive power of the GBRT model is evident in Figure 5, which consists of a scatterplot showing a comparison of actual versus estimated values of SMC (SMC versus SMC*). The evaluation with R^2 = 0.81 and MAE = 0.03 $m^3 \cdot m^{-3}$ suggests that the model can accurately predict the test dataset's SMC. The scatterplots show a slight tendency for underestimation in general but especially for higher SMC values.

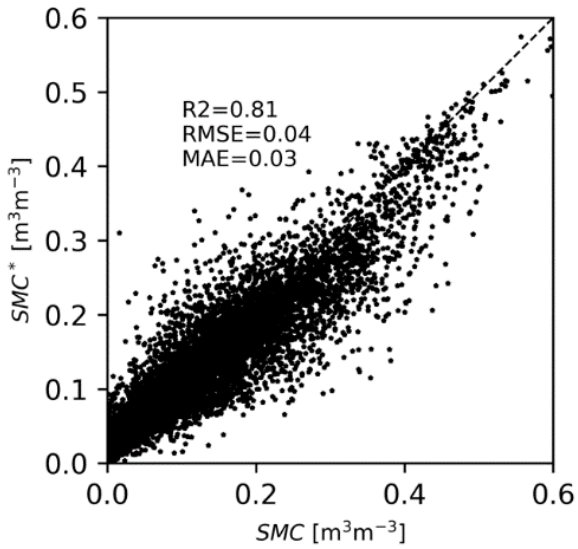

Figure 5. Scatterplot showing the correlation between the estimated SMC* and the true SMC, based on the independent test set.

4.2. Assessment of the Temporal Accuracy

Validation and test scores presented above represent the overall prediction accuracy, combining spatial and temporal errors. The results presented in this section were based on a separate analysis of temporal and spatial variabilities to estimate the associated error components. Table 8 reports the median scores obtained by calculating R^2, RMSE, and MAE (the Spearman correlation ρ, the Pearson correlation R, and the Kling-Gupta Efficiency KGE [89] were included in the table as well to make the results comparable to other SMC products and the comparison in Section 4.5) for each location contained in the test dataset separately, in order to reflect the sensitivity of predicted SMC to temporal variations. The results were averaged over all locations to estimate the overall temporal error and grouped by the land-cover classes. Figure 6 visualizes the same analysis with box and whisker plots, which describe the distribution of RMSE and R^2

Table 8. Median temporal R^2, ρ, R, KGE, RMSE, and MAE overall and grouped by the available land-cover classes.

	Median R^2	Median ρ	Median R	Median KGE	Median RMSE [$m^3 \cdot m^{-3}$]	Median MAE [$m^3 \cdot m^{-3}$]
Overall	0.476	0.702	0.756	0.510	0.037	0.029
Shrubs (20)	0.360	0.650	0.698	0.396	0.033	0.026
Herbaceous vegetation (30)	0.426	0.679	0.742	0.519	0.040	0.033
Cropland (40)	0.518	0.732	0.765	0.542	0.033	0.027
Bare/sparse vegetation (60)	0.258	0.443	0.600	0.240	0.027	0.022
Open forest, evergreen (121)	0.486	0.515	0.756	0.565	0.057	0.045
Open forest, unknown (126)	0.489	0.713	0.750	0.499	0.042	0.035

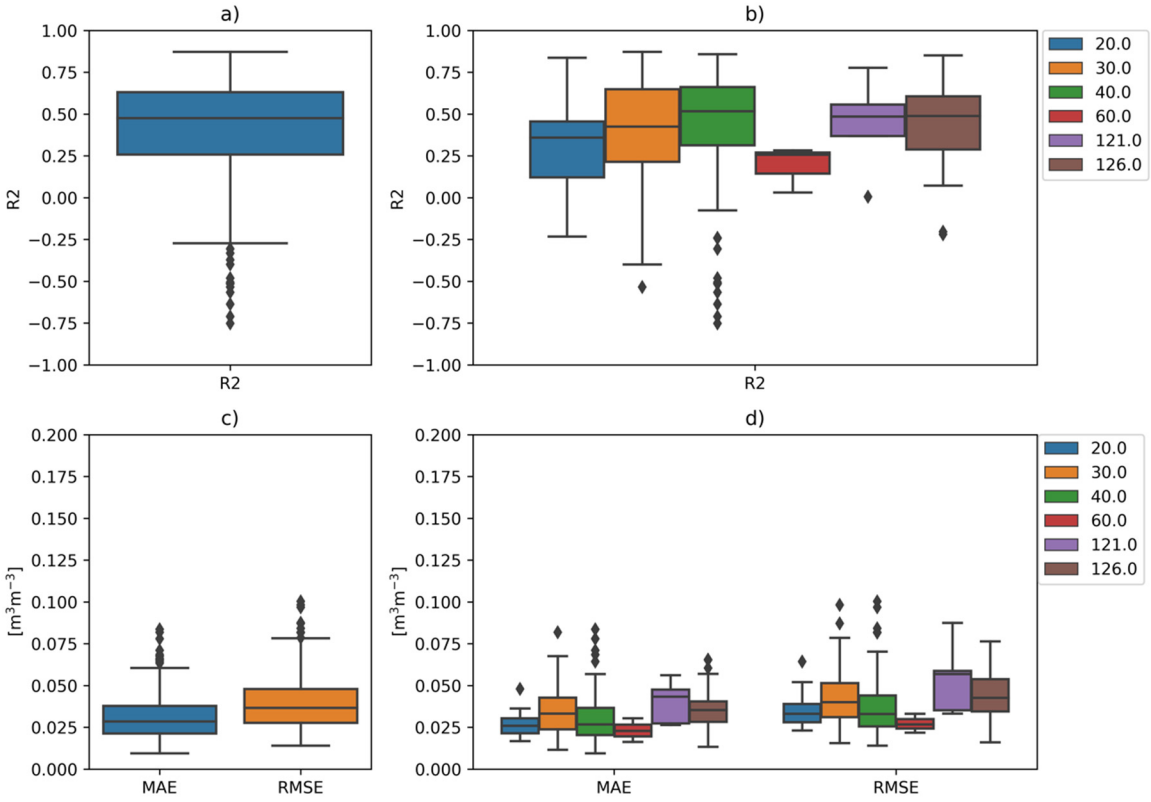

Figure 6. Boxplot summarizing the sensitivity of predicted SMC to temporal dynamics: (**a**) temporal R2, averaged over all locations in the test set; (**b**) temporal R2, averaged over groups of locations belonging to the same land-cover class; (**c**) average accuracy over all locations contained in the test set in terms of MAE and RMSE; (**d**) average accuracy in terms of MAE and RMSE grouped by land-cover class.

With a median of 0.476, the average temporal correlation is significantly lower than the overall value (Figure 5), which hints at a reduced sensitivity to temporal variations compared to spatial variations. The plot in Figure 6b groups the same results by land-cover classes (see Table 3 for an overview of class labels). Locations belonging to the herbaceous vegetation or cropland class (30 or 40) or the two open-forest lasses (121 and 126) show similar R^2 values. However, the quartile range for herbaceous vegetation is large. The average temporal correlation was negatively impacted by the significantly worse results for the class bare/sparse vegetation (60). The boxplots in Figure 6c,d present the same analysis results with respect to MAE and RMSE. Overall, the errors are low, and the differences between the land-cover classes were less distinct compared to those in the analysis based on R^2. It is interesting to note that bare/sparse vegetation performed better, in terms of the error, than the other classes, even though it showed relatively low R^2 values. The low correlation in combination with a low error can be explained by the also low average SMC

of 0.07 m^3m^{-3} for this land-cover class, which indicates that it belongs to an area associated with an arid climate. In [90], Morrison and Wagner demonstrated that the relationship between SMC and radar backscatter in such areas is fundamentally different, caused by the strong effect of surface roughness and the response to subsurface features. Furthermore, the analysis showed that the forest classes have a higher error than vegetation or cropland, even though the results were similar in terms of R^2.

By analyzing the temporally averaged SMC, the discrepancy between the temporal accuracies presented above and the overall accuracy can be explained. The correlation between the average of all actual and estimated SMC values of the same location (Figure 7) shows a very accurate estimation of the static spatial differences (R^2 = 0.97, RMSE = 0.01 m^3m^{-3}, MAE = 0.01 m^3m^{-3}), which contributes to the high overall accuracy of the results.

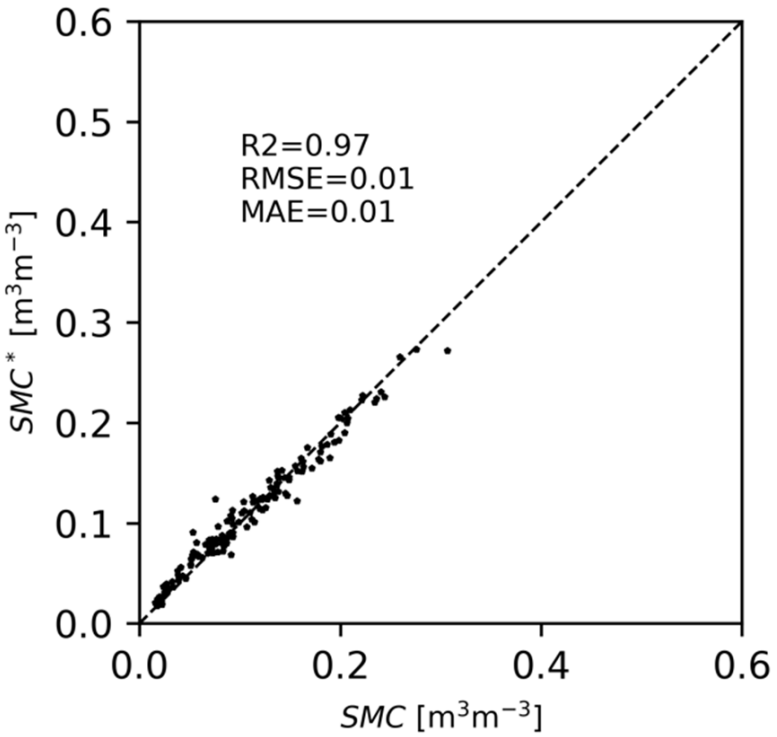

Figure 7. The scatterplot shows the prediction's sensitivity to SMC's static spatial variability, based on comparing the median predicted SMC (SMC*) with the median true SMC.

4.3. Feature Importances

These results were achieved by building the GBRT model based on an automatic feature selection from a large pool of 62 candidate features. Table 9 shows a list of the 16 optimal features. Omitting any of these features would result in a degradation of the predictive power of the estimation model. In the table's far-right column, the relative feature importance is reported, which is a direct output of the GBRT algorithm. It is based on the number of times a feature is used to decide in one of the tree nodes. The importance scores of all features sum up to 1. Figure 8 enables a better interpretation of this relative contribution based on the accumulation by the sources of the features. It emphasizes the significant contribution of optical data and auxiliary datasets and that, despite the proven physical relationship between SMC and S1, estimations could not be based on backscatter intensities without these.

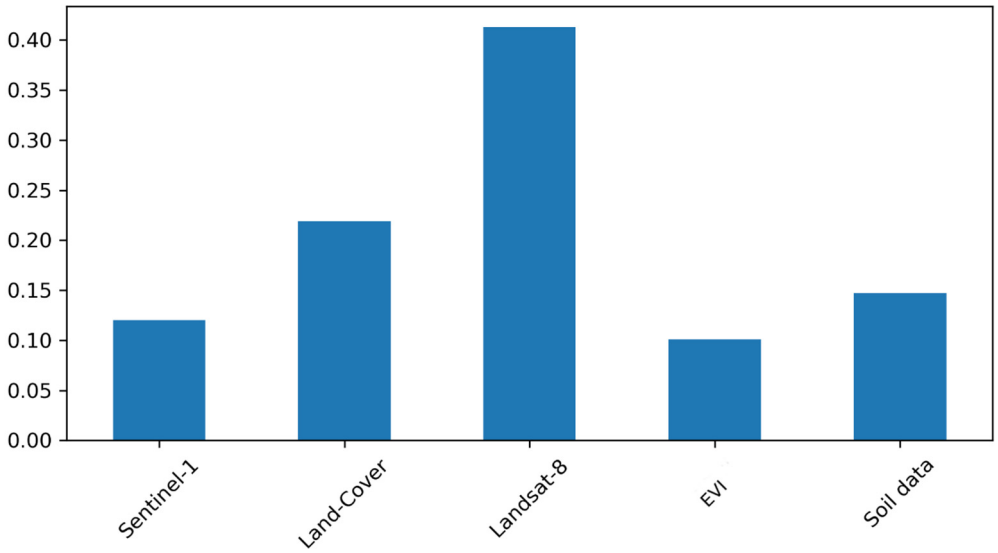

Figure 8. Feature importance aggregated by the original data set or sensor type.

Table 9. Overview of the automatically features ordered by their importance for the retrieval model.

Variable Name	Variable Type	Importance
In situ surface SMC from the ISMN covering the topmost layer of soil (0 to 5 cm)	target	-
Temporal median of L8, band 10	feature	0.134
Percentage of cropland	feature	0.095
Temporal median of S1 $\gamma^0_{VV}vol.$	feature	0.078
Percentage of grassland	feature	0.066
Temporal median of EVI	feature	0.060
Percentage of moss and lichen	feature	0.058
L8, band 4	feature	0.057
EVI	feature	0.056
Temporal median of L8, band 5	feature	0.051
L8, band 7	feature	0.050
Temporal median of L8, band 1	feature	0.047
Soil bulk density	feature	0.047
No. of days between L8 and S1 acquisitions	feature	0.044
S1 $\gamma^0_{VV}vol.$	feature	0.043
L8, band 5	feature	0.041
Percentage of sandy soils	feature	0.040
L8, band 3	feature	0.032

Similar results were achieved in other studies [30], which showed the high impact of auxiliary and optical data on SAR-based retrieval of SMC. The authors demonstrated that this phenomenon is due to the combination of an indirect relationship between vegetation properties and SMC and their high impact on the backscatter intensities. Research presented in several other articles [29,91,92] confirmed these findings by analyzing the impact of the vegetation phenology and surface roughness on Sentinel-1 backscatter intensities.

4.4. Training Performance

Figure 9 presents further insights into the estimation model performance. The learning curve compares the training and cross-validation scores which are dependent on the size of the training dataset. It shows that a large dataset is necessary to increase the generalization capabilities of the retrieval model. In this case, the cross-validation score curve still shows a positive trend, even with the maximum number of samples; therefore, an increased estimation accuracy could be expected with a more extensive training dataset. The model's scalability shows that the cost (in terms of computational time) of increasing the number

of training samples is linear. However, the gain, in terms of increased score, shows a nonlinear behavior. Due to the linear relationship between the number of training samples and computational effort, this curve shows identical behavior as the curve given by the cross-validation score.

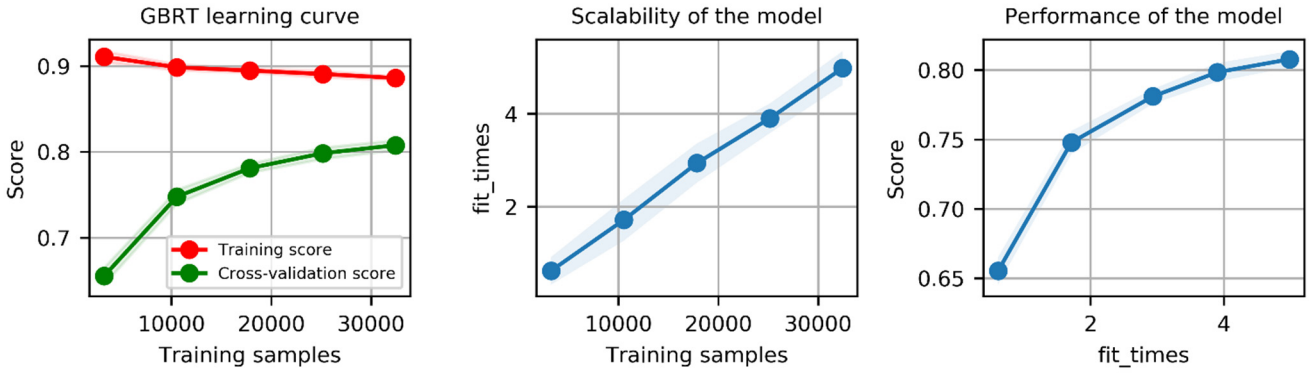

Figure 9. GBRT model performance in terms of the learning curve, the model scalability, and model performance.

4.5. Comparison to Established Methods and Other Experimental Results

This section compares results presented in this article to those from established SMC products, as well as results in recently published articles. Certain limitations apply: quantification of the differences is difficult because the underlying SMC products are fundamentally different, for example, in terms of spatial resolution or the characteristics of the underlying remote-sensing data, and the validation approaches apply different methods and reference datasets. The numbers presented in Table 10 are, therefore, intended for a qualitative assessment.

Table 10. Comparison of correlations between actual and estimated SMC, based on the method introduced in this article (called PYSMM in the table) with validation results of other established soil moisture products and experimental results.

	Median Temporal R^2	Median Temporal ρ	Median Temporal R	Overall R^2	Reference
PYSMM	0.476	0.702	0.756	0.81	
CSSM	-	0.315	-	-	[93]
SMAP	-	-	0.752	-	[94]
RFS1				0.682	[40]

One of the freely available operational SMC products with the highest spatial resolution is the Copernicus Surface Soil Moisture (CSSM) product [19], based on Sentinel-1 data. It provides a mapping of the SMC with a spatial resolution of 1 km. A validation report [93] presents an assessment of the temporal correlation between estimated SMC and measurements from several in situ networks, which are part of the ISMN. The result was a median temporal correlation of ρ = 0.315. Like in the analysis presented in Section 4.2, the range of ρ values for the individual sites is significant.

The Soil Moisture Active Passive (SMAP) mission [15] provides several SMC products derived from passive only and a combination of active and passive microwave data. The L2SMAP product is providing SMC data at a spatial resolution of 9 km. In [94], Colliander et al. performed the validation of several SMAP products based on the so-called core validation sites, which are in situ networks established explicitly to validate SMAP. In terms of the Pearson correlation coefficient (R), the reported median temporal correlation for the L2SMAP product was R = 0.752.

Chatterjee et al. [40] presented a study with a similar aim, as presented in this article. The authors tested several ML approaches to estimate SMC for the continental USA, also based on Sentinel-1 data. Training and validation were performed based on in situ

measurements from the USCRN network. The overall correlation between actual and estimated SMC, derived with a Random-Forest based model (RFS1), was $R^2 = 0.682$.

5. Conclusions

This study introduced an approach to estimate SMC at a high spatial resolution on a quasi-global scale. The novelty of this approach is the application of a data-driven model in a large-scale context. One of its strengths is that the mapping of SMC is cloud-based, which means that newly available reference data can be easily integrated, and the model retrained without the necessity to process further satellite data. A Python package, called PYSMM [88], for the online retrieval of SMC and a demonstrator dataset was developed to supplement this study. The validation demonstrated that, within certain boundary conditions, an overall high accuracy could be achieved. Compared to the performance of available SMC products as well as similar studies, the results are promising. The overall accuracy fulfils SMC monitoring requirements set by the Global Observing System (GCOS) in [95], which specifies an overall retrieval error of fewer than 0.04 m^3m^{-3}. Certain main limitations do apply: (1) the irregular distribution of samples within the feature space leads to variable accuracies (as the results in Section 4.2 show based on the example of the land-cover class bare/sparse vegetation); (2) the approach is limited to low vegetation density areas. Based on the CGLS classification, this reduces the mappable area, for example, in Europe, to about 55% of the total land area and about 66% in the USA, and as low as 15% in Indonesia, which is further reduced by masking high NDVI values, frozen soil, or snow; (3) the training dataset covers only some of the global climate zones (Figure 1). Especially point 1 could be tackled in future research by identifying and targeting the low sampled areas of the feature space. Future work should also focus on extending the model to incorporate other remote-sensing sensors like Sentinel-2 and the collection of reference data for currently missing climate zones.

Two studies achieved promising results based on SMC estimations derived with PYSMM, demonstrating spatial and temporal mapping potential. Lei et al. [96] showed how SMC maps could be assimilated into a hydrological model to improve the spatial details of the model simulations, and Greifeneder et al. [97] combined time-series of estimated SMC with GLDAS soil moisture climatologies to derive temporal anomalies.

Supplementary Materials: The comparison of GBRT with three further ML algorithms is provided in S1: Algorithm Comparison, available online at: http://doi.org/10.5281/zenodo.4742678; S2, the PYMM source code is available online at http://doi.org/10.5281/zenodo.4552813 The documentation (S3) is available directly here: https://pysmm.readthedocs.io/en/latest/. Two SMC demonstrator data sets (S4) can be viewed and downloaded through a Google Earth Engine App (https://felixgreifeneder.users.earthengine.app/view/sm-explorer).

Author Contributions: Conceptualization, F.G.; Methodology, F.G.; Supervision, C.N. and W.W.; Writing—original draft, F.G.; Writing—review & editing, C.N. and W.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially funded by the Horizon 2020 project "Ecopotential—Improving Future Ecosystem Benefits through Earth Observation, which has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 641762) and the European Fund for Regional Development project "DPS4ESLAB".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available in a publicly accessible repository that does not issue DOIs. Publicly available datasets were analyzed in this study. This data can be found here: https://ismn.tuwien.ac.at and https://developers.google.com/earth-engine/datasets, accessed on 28 February 2020.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Bras, R.L. Complexity and organization in hydrology: A personal view. *Water Resour. Res.* **2015**, *51*, 6532–6548. [CrossRef]
2. Liu, D.; Wang, G.; Mei, R.; Yu, Z.; Yu, M. Impact of initial soil moisture anomalies on climate mean and extremes over Asia. *J. Geophys. Res. Atmos.* **2014**, *119*, 529–545. [CrossRef]
3. Ulaby, F.T.; Batlivala, P.P.; Dobson, M.C. Microwave Backscatter Dependence on Surface Roughness, Soil Moisture, and Soil Texture: Part I-Bare Soil. *IEEE Trans. Geosci. Electron.* **1978**, *16*, 286–295. [CrossRef]
4. Zhang, D.; Zhou, G. Estimation of Soil Moisture from Optical and Thermal Remote Sensing: A Review. *Sensors* **2016**, *16*, 1308. [CrossRef]
5. Fung, A.K.; Li, Z.; Chen, K.S. Backscatterin from randomly rough dielectric surface. *IEEE Trans. Geosci. Remote Sens.* **1992**, *30*, 356–369. [CrossRef]
6. Oh, Y.; Sarabandi, K.; Ulaby, F.T. An empirical model and an inversion technique for radar scattering from bare soil surfaces. *IEEE Trans. Geosci. Remote Sens.* **1992**, *30*, 370–381. [CrossRef]
7. Dubois, P.; Van Zyl, J.; Engman, T. Measuring soil moisture with imaging radars. *IEEE Trans. Geosci. Remote Sens.* **1995**, *33*, 916–926. [CrossRef]
8. Wagner, W.; Lemoine, G.; Borgeaud, M.; Member, S.; Rott, H. A Study of Vegetation Cover Effects on ERS Scatterometer Data. *Geosci. Remote Sens. IEEE Trans.* **1999**, *37*, 938–948. [CrossRef]
9. Ångström, A. The Albedo of Various Surfaces of Ground. *Geogr. Ann.* **1925**, *7*, 323–342. [CrossRef]
10. Qin, J.; Yang, K.; Lu, N.; Chen, Y.; Zhao, L.; Han, M. Spatial upscaling of in-situ soil moisture measurements based on MODIS-derived apparent thermal inertia. *Remote Sens. Environ.* **2013**, *138*, 1–9. [CrossRef]
11. Schmugge, T. Remote Sensing of Surface Soil Moisture. *J. Appl. Meteorol. Climatol.* **1978**, *17*, 1549–1557. [CrossRef]
12. Chang, T.-Y.; Wang, Y.-C.; Feng, C.-C.; Ziegler, A.D.; Giambelluca, T.W.; Liou, Y.-A. Estimation of Root Zone Soil Moisture Using Apparent Thermal Inertia With MODIS Imagery Over a Tropical Catchment in Northern Thailand. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2012**, *5*, 752–761. [CrossRef]
13. Price, J. Using spatial context in satellite data to infer regional scale evapotranspiration. *IEEE Trans. Geosci. Remote Sens.* **1990**, *28*, 940–948. [CrossRef]
14. Berger, M.; Camps, A.; Font, J.; Kerr, Y.; Miller, J.; Johannessen, J.A.; Boutin, J.; Drinkwater, M.R.; Skou, N.; Floury, N.; et al. Measuring ocean salinity with ESA's SMOS mission—Advancing the science. *ESA Bull. Eur. Space Agency* **2002**, *111*, 113–121.
15. 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]
16. Naeimi, V.; Scipal, K.; Bartalis, Z.; Hasenauer, S.; Wagner, W. An Improved Soil Moisture Retrieval Algorithm for ERS and METOP Scatterometer Observations. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 1999–2013. [CrossRef]
17. Bauer-Marschallinger, B.; Paulik, C.; Hochstöger, S.; Mistelbauer, T.; Modanesi, S.; Ciabatta, L.; Massari, C.; Brocca, L.; Wagner, W. Soil Moisture from Fusion of Scatterometer and SAR: Closing the Scale Gap with Temporal Filtering. *Remote Sens.* **2018**, *10*, 1030. [CrossRef]
18. Hornacek, M.; Wagner, W.; Sabel, D.; Truong, H.-L.; Snoeij, P.; Hahmann, T.; Diedrich, E.; Doubkova, M. Potential for High Resolution Systematic Global Surface Soil Moisture Retrieval via Change Detection Using Sentinel-1. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2012**, *5*, 1303–1311. [CrossRef]
19. Bauer-Marschallinger, B.; Freeman, V.; Cao, S.; Paulik, C.; Schaufler, S.; Stachl, T.; Modanesi, S.; Massari, C.; Ciabatta, L.; Brocca, L.; et al. Toward Global Soil Moisture Monitoring With Sentinel-1: Harnessing Assets and Overcoming Obstacles. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 520–539. [CrossRef]
20. Ali, I.; Greifeneder, F.; Stamenkovic, J.; Neumann, M.; Notarnicola, C. Review of Machine Learning Approaches for Biomass and Soil Moisture Retrievals from Remote Sensing Data. *Remote Sens.* **2015**, *7*, 16398–16421. [CrossRef]
21. Bhuiyan, M.A.E.; Anagnostou, E.N.; Kirstetter, P.-E. A Nonparametric Statistical Technique for Modeling Overland TMI (2A12) Rainfall Retrieval Error. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1898–1902. [CrossRef]
22. Tyralis, H.; Papacharalampous, G.; Burnetas, A.; Langousis, A. Hydrological post-processing using stacked generalization of quantile regression algorithms: Large-scale application over CONUS. *J. Hydrol.* **2019**, *577*, 123957. [CrossRef]
23. Derin, Y.; Bhuiyan, A.E.; Anagnostou, E.; Kalogiros, J.; Anagnostou, M.N. Modeling Level 2 Passive Microwave Precipitation Retrieval Error Over Complex Terrain Using a Nonparametric Statistical Technique. *IEEE Trans. Geosci. Remote Sens.* **2020**, 1–12. [CrossRef]
24. Ahn, K.-H.; Palmer, R. Regional flood frequency analysis using spatial proximity and basin characteristics: Quantile regression vs. parameter regression technique. *J. Hydrol.* **2016**, *540*, 515–526. [CrossRef]
25. Aboutalebi, M.; Allen, N.; Torres-Rua, A.F.; Coopmans, C. Estimation of soil moisture at different soil levels using machine learning techniques and unmanned aerial vehicle (UAV) multispectral imagery. In *Autonomous Air and Ground Sensing Systems for Agricultural Optimization and Phenotyping IV*; International Society for Optical Engineering: Baltimore, MD, USA, 2019; p. 26. [CrossRef]

26. Moosavi, V.; Talebi, A.; Mokhtari, M.H.; Hadian, M.R. Estimation of spatially enhanced soil moisture combining remote sensing and artificial intelligence approaches. *Int. J. Remote Sens.* **2016**, *37*, 5605–5631. [CrossRef]
27. Srivastava, P.K.; Han, D.; Ramirez, M.R.; Islam, T. Machine Learning Techniques for Downscaling SMOS Satellite Soil Moisture Using MODIS Land Surface Temperature for Hydrological Application. *Water Resour. Manag.* **2013**, *27*, 3127–3144. [CrossRef]
28. Li, Y.; Yan, S.; Chen, N.; Gong, J. Performance Evaluation of a Neural Network Model and Two Empirical Models for Estimating Soil Moisture Based on Sentinel-1 SAR Data. *Prog. Electromagn. Res. C* **2020**, *105*, 85–99. [CrossRef]
29. Paloscia, S.; Pettinato, S.; Santi, E.; Notarnicola, C.; Pasolli, L.; Reppucci, A. Soil moisture mapping using Sentinel-1 images: Algorithm and preliminary validation. *Remote Sens. Environ.* **2013**, *134*, 234–248. [CrossRef]
30. Pasolli, L.; Notarnicola, C.; Bertoldi, G.; Bruzzone, L.; Remelgado, R.; Greifeneder, F.; Niedrist, G.; Della Chiesa, S.; Tappeiner, U.; Zebisch, M. Estimation of Soil Moisture in Mountain Areas Using SVR Technique Applied to Multiscale Active Radar Images at C-Band. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 262–283. [CrossRef]
31. Santi, E.; Paloscia, S.; Pettinato, S.; Fontanelli, G. A prototype ann based algorithm for the soil moisture retrieval from l- band in view of the incoming SMAP mission. In *2014 13th Specialist Meeting on Microwave Radiometry and Remote Sensing of the Environment (MicroRad)*; IEEE: Pasadena, CA, USA, 2014; pp. 5–9. [CrossRef]
32. Stamenkovic, J.; Guerriero, L.; Ferrazzoli, P.; Notarnicola, C.; Greifeneder, F.; Thiran, J.-P. Soil Moisture Estimation by SAR in Alpine Fields Using Gaussian Process Regressor Trained by Model Simulations. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 4899–4912. [CrossRef]
33. Kolassa, J.; Reichle, R.H.; Liu, Q.; Alemohammad, S.H.; Gentine, P.; Aida, K.; Asanuma, J.; Bircher, S.; Caldwell, T.; Colliander, A.; et al. Estimating surface soil moisture from SMAP observations using a Neural Network technique. *Remote Sens. Environ.* **2018**, *204*, 43–59. [CrossRef]
34. Liu, Y.; Qian, J.; Yue, H. Combined Sentinel-1A With Sentinel-2A to Estimate Soil Moisture in Farmland. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 1292–1310. [CrossRef]
35. Bhuiyan, M.A.E.; Yang, F.; Biswas, N.K.; Rahat, S.H.; Neelam, T.J. Machine Learning-Based Error Modeling to Improve GPM IMERG Precipitation Product over the Brahmaputra River Basin. *Forecasting* **2020**, *2*, 248–266. [CrossRef]
36. Hird, J.N.; DeLancey, E.R.; McDermid, G.J.; Kariyeva, J. Google Earth Engine, Open-Access Satellite Data, and Machine Learning in Support of Large-Area Probabilistic Wetland Mapping. *Remote Sens.* **2017**, *9*, 1315. [CrossRef]
37. Cho, E.; Jacobs, J.M.; Jia, X.; Kraatz, S. Identifying Subsurface Drainage using Satellite Big Data and Machine Learning via Google Earth Engine. *Water Resour. Res.* **2019**, *55*, 8028–8045. [CrossRef]
38. Traganos, D.; Aggarwal, B.; Poursanidis, D.; Topouzelis, K.; Chrysoulakis, N.; Reinartz, P. Towards Global-Scale Seagrass Mapping and Monitoring Using Sentinel-2 on Google Earth Engine: The Case Study of the Aegean and Ionian Seas. *Remote Sens.* **2018**, *10*, 1227. [CrossRef]
39. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **2017**, *202*, 18–27. [CrossRef]
40. Chatterjee, S.; Huang, J.; Hartemink, A.E. Establishing an Empirical Model for Surface Soil Moisture Retrieval at the U.S. Climate Reference Network Using Sentinel-1 Backscatter and Ancillary Data. *Remote Sens.* **2020**, *12*, 1242. [CrossRef]
41. Dorigo, W.; Wagner, W.; Hohensinn, R.; Hahn, S.; Paulik, C.; Xaver, A.; Gruber, A.; Drusch, M.; Mecklenburg, S.; Van Oevelen, P.; et al. The International Soil Moisture Network: A data hosting facility for global in situ soil moisture measurements. *Hydrol. Earth Syst. Sci.* **2011**, *15*, 1675–1698. [CrossRef]
42. Dorigo, W.A.; Xaver, A.; Vreugdenhil, M.; Gruber, A.; Hegyiová, A.; Sanchis-Dufau, A.D.; Zamojski, D.; Cordes, C.; Wagner, W.; Drusch, M. Global Automated Quality Control of In Situ Soil Moisture Data from the International Soil Moisture Network. *Vadose Zone J.* **2013**, *12*, vzj2012.0097. [CrossRef]
43. Paulik, C.; Plocon, A.; Hahn, S.; Mistelbauer, T.; Reimer, C. TUW-GEO/pytesmo. *Pytesmo* **2017**. [CrossRef]
44. Beck, H.E.; Zimmermann, N.E.; McVicar, T.R.; Vergopolan, N.; Berg, A.; Wood, E.F. Present and future Köppen-Geiger climate classification maps at 1-km resolution. *Sci. Data* **2018**, *5*, 180214. [CrossRef] [PubMed]
45. Bell, J.E.; Palecki, M.A.; Baker, C.B.; Collins, W.G.; Lawrimore, J.H.; Leeper, R.; Hall, M.E.; Kochendorfer, J.; Meyers, T.P.; Wilson, T.; et al. U.S. Climate Reference Network Soil Moisture and Temperature Observations. *J. Hydrometeorol.* **2013**, *14*, 977–988. [CrossRef]
46. Smith, A.B.; Walker, J.P.; Western, A.W.; Young, R.I.; Ellett, K.M.; Pipunic, R.C.; Grayson, R.B.; Siriwardena, L.; Chiew, F.H.S.; Richter, H.G. The Murrumbidgee soil moisture monitoring network data set. *Water Resour. Res.* **2012**, *48*, 1–6. [CrossRef]
47. Young, R.; Walker, J.P.; Yeoh, N.; Smith, A.; Ellett, K.M.; Merlin, O.; Western, A. Soil Moisture and Meteorological Observations from the Murrumbidgee Catchment. 2008. Available online: http://www.oznet.org.au/documentation/Soil_Moisture_Meteorological_Observation_of_Murrumbidgee_Catchment.pdf (accessed on 28 February 2020).
48. Larson, K.M.; Small, E.E.; Gutmann, E.D.; Bilich, A.L.; Braun, J.J.; Zavorotny, V.U. Use of GPS receivers as a soil moisture network for water cycle studies. *Geophys. Res. Lett.* **2008**, *35*, 24405. [CrossRef]
49. Moghaddam, M.; Entekhabi, D.; Goykhman, Y.; Li, K.; Liu, M.; Mahajan, A.; Nayyar, A.; Shuman, D.; Teneketzis, D. A Wireless Soil Moisture Smart Sensor Web Using Physics-Based Optimal Control: Concept and Initial Demonstrations. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2010**, *3*, 522–535. [CrossRef]
50. Moghaddam, M.; Silva, A.; Clewley, D.; Akbar, R.; Hussaini, S.A.; Whitcomb, J.; Devarakonda, R.; Shrestha, R.; Cook, R.B.; Prakash, G.; et al. *Soil Moisture Profiles and Temperature Data from SoilSCAPE Sites, USA*; ORNL DAAC: Oak Ridge, TN, USA, 2017.

51. Bircher, S.; Skou, N.; Jensen, K.H.; Walker, J.P.; Rasmussen, L. A soil moisture and temperature network for SMOS validation in Western Denmark. *Hydrol. Earth Syst. Sci.* **2012**, *16*, 1445–1463. [CrossRef]
52. Albergel, C.; Rüdiger, C.; Pellarin, T.; Calvet, J.-C.; Fritz, N.; Froissard, F.; Suquia, D.; Petitpa, A.; Piguet, B.; Martin, E. From near-surface to root-zone soil moisture using an exponential filter: An assessment of the method based on in-situ observations and model simulations. *Hydrol. Earth Syst. Sci.* **2008**, *12*, 1323–1337. [CrossRef]
53. Calvet, J.-C.; Fritz, N.; Froissard, F.; Suquia, D.; Petitpa, A.; Piguet, B. In situ soil moisture observations for the CAL/VAL of SMOS: The SMOSMANIA network. In *2007 IEEE International Geoscience and Remote Sensing Symposium*; IEEE: Barcelona, Spain, 2007; pp. 1196–1199. [CrossRef]
54. Ojo, E.R.; Bullock, P.R.; L'Heureux, J.; Powers, J.; McNairn, H.; Pacheco, A. Calibration and evaluation of a frequency domain reflectometry sensor for real-time soil moisture monitoring. *Vadose Zone J.* **2015**, *14*, 1–15. [CrossRef]
55. Osenga, E.C.; Arnott, J.C.; Endsley, K.A.; Katzenberger, J.W. Bioclimatic and Soil Moisture Monitoring Across Elevation in a Mountain Watershed: Opportunities for Research and Resource Management. *Water Resour. Res.* **2019**, *55*, 2493–2503. [CrossRef]
56. Zacharias, S.; Bogena, H.; Samaniego, L.; Mauder, M.; Fuß, R.; Pütz, T.; Frenzel, M.; Schwank, M.; Baessler, C.; Butterbach-Bahl, K.; et al. A Network of Terrestrial Environmental Observatories in Germany. *Vadose Zone J.* **2011**, *10*, 955–973. [CrossRef]
57. Tagesson, T.; Fensholt, R.; Guiro, I.; Rasmussen, M.O.; Huber, S.; Mbow, C.; Garcia, M.; Horion, S.; Sandholt, I.; Holm-Rasmussen, B.; et al. Ecosystem properties of semiarid savanna grassland in West Africa and its relationship with environmental variability. *Glob. Chang. Biol.* **2015**, *21*, 250–264. [CrossRef]
58. Wagner, W. Big Data infrastructures for processing Sentinel data. In *Photogramm. Week*; University of Stuttgart: Stuttgart, Germany, 2015; pp. 93–104.
59. European Space Agency. Sentinel-1 User Handbook. European Space Agency, European Comission. 2013. Available online: https://sentinel.esa.int (accessed on 16 December 2020).
60. Buchhorn, M.; Smets, B.; Bertels, L.; Lesiv, M.; Tsendbazar, N.-E.; Herold, M.; Fritz, S. Copernicus Global Land Service: Land Cover 100m: Collection 3: Epoch 2015: Globe. *Zenodo* **2020**. [CrossRef]
61. Rodell, M.; Houser, P.R.; Jambor, U.; Gottschalck, J.; Mitchell, K.; Meng, C.-J.; Arsenault, K.; Cosgrove, B.; Radakovich, J.; Bosilovich, M.; et al. The Global Land Data Assimilation System. *Bull. Am. Meteorol. Soc.* **2004**, *85*, 381–394. [CrossRef]
62. Chander, G.; Markham, B.L.; Helder, D.L. Summary of current radiometric calibration coefficients for Landsat MSS, TM, ETM+, and EO-1 ALI sensors. *Remote Sens. Environ.* **2009**, *113*, 893–903. [CrossRef]
63. Didan, K. *MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V006, V006 ed.*; NASA EOSDIS LP DAAC: Sioux Falls, SD, USA, 2015. [CrossRef]
64. Hengl, T. Soil texture classes (USDA system) for 6 soil depths (0, 10, 30, 60, 100 and 200 cm) at 250 m. *Zenodo* **2018**. [CrossRef]
65. Hengl, T. Soil bulk density (fine earth) 10 x kg/m-cubic at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution. *Zenodo* **2018**. [CrossRef]
66. Hengl, T. Clay content in % (kg/kg) at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution. *Zenodo* **2018**. [CrossRef]
67. Hengl, T. Sand content in % (kg/kg) at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution. *Zenodo* **2018**. [CrossRef]
68. Google. Google Earth Engine: Sentinel-1 Pre-Processing. 2016. Available online: https://developers.google.com/earth-engine/sentinel1 (accessed on 18 December 2017).
69. Quegan, S.; Le Toan, T.; Yu, J.; Ribbes, F.; Floury, N. Multitemporal ERS SAR analysis applied to forest mapping. *IEEE Trans. Geosci. Remote Sens.* **2000**, *38*, 741–753. [CrossRef]
70. Vollrath, A.; Mullissa, A.; Reiche, J. Angular-Based Radiometric Slope Correction for Sentinel-1 on Google Earth Engine. *Remote Sens.* **2020**, *12*, 1867. [CrossRef]
71. Hoekman, D.H. Radar Remote Sensing Data for Applications in Forestry. Dissertation, Internally Prepared, Laboratory of Geo-Information Science and Remote Sensing, Wageningen University, Wageningen, The Netherlands. 1990. Available online: https://library.wur.nl/WebQuery/wurpubs/12860 (accessed on 12 January 2021).
72. Ulander, L. Radiometric slope correction of synthetic-aperture radar images. *IEEE Trans. Geosci. Remote Sens.* **1996**, *34*, 1115–1122. [CrossRef]
73. Farr, T.G.; Rosen, P.A.; Caro, E.; Crippen, R.; Duren, R.; Hensley, S.; Kobrick, M.; Paller, M.; Rodriguez, E.; Roth, L.; et al. The Shuttle Radar Topography Mission. *Rev. Geophys.* **2007**, *45*, 1–33. [CrossRef]
74. Beaudoin, A.; Le Toan, T.; Gwyn, Q. SAR observations and modeling of the C-band backscatter variability due to multiscale geometry and soil moisture. *IEEE Trans. Geosci. Remote Sens.* **1990**, *28*, 886–895. [CrossRef]
75. Wagner, W.; Lemoine, G.; Rott, H. A Method for Estimating Soil Moisture from ERS Scatterometer and Soil Data. *Remote Sens. Environ.* **1999**, *70*, 191–207. [CrossRef]
76. Verrelst, J.; Muñoz, J.; Alonso, L.; Delegido, J.; Rivera, J.P.; Camps-Valls, G.; Moreno, J. Machine learning regression algorithms for biophysical parameter retrieval: Opportunities for Sentinel-2 and -3. *Remote Sens. Environ.* **2012**, *118*, 127–139. [CrossRef]
77. Upreti, D.; Huang, W.; Kong, W.; Pascucci, S.; Pignatti, S.; Zhou, X.; Ye, H.; Casa, R. A Comparison of Hybrid Machine Learning Algorithms for the Retrieval of Wheat Biophysical Variables from Sentinel-2. *Remote Sens.* **2019**, *11*, 481. [CrossRef]
78. Duffy, N.; Helmbold, D. Boosting Methods for Regression. *Mach. Learn.* **2002**, *47*, 153–200. [CrossRef]

79. Elith, J.; Leathwick, J.R.; Hastie, T. A working guide to boosted regression trees. *J. Anim. Ecol.* **2008**, *77*, 802–813. [CrossRef] [PubMed]
80. Pham, T.D.; Yokoya, N.; Xia, J.; Ha, N.T.; Le, N.N.; Nguyen, T.T.T.; Dao, T.H.; Vu, T.T.P.; Takeuchi, W. Comparison of Machine Learning Methods for Estimating Mangrove Above-Ground Biomass Using Multiple Source Remote Sensing Data in the Red River Delta Biosphere Reserve, Vietnam. *Remote Sens.* **2020**, *12*, 1334. [CrossRef]
81. Hrisko, J.; Ramamurthy, P.; Gonzalez, J.E. Estimating heat storage in urban areas using multispectral satellite data and machine learning. *Remote Sens. Environ.* **2021**, *252*, 112125. [CrossRef]
82. Wang, Y.; Jiang, B.; Liang, S.; Wang, D.; He, T.; Wang, Q.; Zhao, X.; Xu, J. Surface Shortwave Net Radiation Estimation from Landsat TM/ETM+ Data Using Four Machine Learning Algorithms. *Remote Sens.* **2019**, *11*, 2847. [CrossRef]
83. Barzilai, J.; Borwein, J.M. Two-Point Step Size Gradient Methods. *Ima J. Numer. Anal.* **1988**, *8*, 141–148. [CrossRef]
84. Armijo, L. Minimization of functions having Lipschitz continuous first partial derivatives. *Pac. J. Math.* **1966**, *16*, 1–3. [CrossRef]
85. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
86. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
87. Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *Int. Jt. Conf. Artif. Intell.* **1995**, *14*, 1137–1143. [CrossRef]
88. Greifeneder, F. PYSMM. *Zenodo* **2021**. [CrossRef]
89. 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]
90. Morrison, K.; Wagner, W. Explaining Anomalies in SAR and Scatterometer Soil Moisture Retrievals From Dry Soils With Subsurface Scattering. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 2190–2197. [CrossRef]
91. Stendardi, L.; Karlsen, S.R.; Niedrist, G.; Gerdol, R.; Zebisch, M.; Rossi, M.; Notarnicola, C. Exploiting Time Series of Sentinel-1 and Sentinel-2 Imagery to Detect Meadow Phenology in Mountain Regions. *Remote Sens.* **2019**, *11*, 542. [CrossRef]
92. Notarnicola, C.; Angiulli, M.; Posa, F. Use of radar and optical remotely sensed data for soil moisture retrieval over vegetated areas. *IEEE Trans. Geosci. Remote Sens.* **2006**, *44*, 925–935. [CrossRef]
93. Bauer-marschallinger, B.; Schaufler, S.; Navacchi, C. Validation Report Surface Soil Moisture Collection 1KM Version 1. 2018. Available online: https://land.copernicus.eu/global/sites/cgls.vito.be/files/products/CGLOPS1_VR_SSM1km-V1_I1.20.pdf (accessed on 3 November 2020).
94. Colliander, A.; Jackson, T.J.; Bindlish, R.; Chan, S.; Das, N.; Kim, S.B.; Cosh, M.H.; Dunbar, R.S.; Dang, L.; Pashaian, L.; et al. Validation of SMAP surface soil moisture products with core validation sites. *Remote Sens. Environ.* **2017**, *191*, 215–231. [CrossRef]
95. *Systematic Observation Requirements for Satellite-Based Data Products for Climate*; GCOS Secretariat: Geneva, Switzerland, 2011; Available online: https://climate.esa.int/sites/default/files/gcos-154.pdf (accessed on 3 November 2020).
96. Lei, F.; Crow, W.T.; Kustas, W.P.; Dong, J.; Yang, Y.; Knipper, K.R.; Anderson, M.C.; Gao, F.; Notarnicola, C.; Greifeneder, F.; et al. Data assimilation of high-resolution thermal and radar remote sensing retrievals for soil moisture monitoring in a drip-irrigated vineyard. *Remote Sens. Environ.* **2020**, *239*, 111622. [CrossRef]
97. Greifeneder, F.; Khamala, E.; Sendabo, D.; Wagner, W.; Zebisch, M.; Farah, H.; Notarnicola, C. Detection of soil moisture anomalies based on Sentinel-1. *Phys. Chem. Earth Parts ABC* **2018**, 1–8. [CrossRef]