Science of Remote Sensing 10 (2024) 100177

Contents lists available at ScienceDirect

Science of Remote Sensing

journal homepage: www.sciencedirect.com/journal/science-of-remote-sensing

Global soil moisture mapping at 5 km by combining GNSS reflectometry and machine learning in view of HydroGNSS

Emanuele Santi [a,*], Davide Comite [b], Laura Dente [c], Leila Guerriero [c], Nazzareno Pierdicca [b], Maria Paola Clarizia [d], Nicolas Floury [d]

[a] Institute of Applied Physics, National Research Council (IFAC-CNR), Firenze, Italy
[b] Department of Information Engineering, Electronics, Telecommunications, La Sapienza University of Rome, Italy
[c] Dipartimento di Ingegneria Civile e Ingegneria Informatica, Tor Vergata University, Rome, Italy
[d] European Space Agency - European Space Technology and Research Centre ESA-ESTEC, Noordwijk, the Netherlands

ARTICLE INFO

Index Terms:
GNSS reflectometry (GNSS-R)
CyGNSS
Soil moisture
Machine learning

ABSTRACT

The potential of GNSS reflectometry (GNSS-R) for the monitoring of soil and vegetation parameters as soil moisture (SM) and forest aboveground biomass (AGB) has been largely investigated in recent years.

In view of the ESA's HydroGNSS mission, planned to be launched in 2024, this study has explored the possibility to map SM at global scale and relatively high resolution of about 0.05° (corresponding approximately to 5 Km) using GNSS-R observations, by implementing and comparing two retrieval algorithms based on machine learning techniques, namely Artificial Neural Networks (ANN) and Random Forest Regressors (RF). Waiting for HydroGNSS commissioning and operation, the NASA's Cyclone GNSS (CyGNSS) land observations have been considered for this scope. Taking advantage of the versatility of both machine learning techniques, several combinations of input data, including CyGNSS observables and auxiliary information, have been exploited and the role of GNSS-R and auxiliary data has been assessed. Given the lack of global SM data at 0.05° resolution, the following novel strategy has been implemented to establish the training set: as first, training has been carried out at lower resolution by considering as target the SMAP SM on EASE-Grid 36 km. Then the trained algorithms have been applied to CyGNSS data at 0.05° to obtain global SM maps at this resolution. Finally, the SM at 0.05° has been validated against ISMN, to keep training and validation as much independent as possible. The two retrieval techniques exhibited similar accuracies and computational cost, with correlation coefficient R $\simeq$ 0.9 between estimated and target SM computed globally, and RMSE $\simeq$ 0.05 (m^3/m^3). Moreover, the SM maps at 0.05° revealed some finer details and small-scale patterns that are not shown by the original SMAP SM data at 36 km. Regardless of the ML technique applied, this study confirmed the promising potential of GNSS-R for the global monitoring of SM at improved resolution with respect to SM products available from microwave satellite radiometers.

1. Introduction

The use of Global Navigation Satellite System Reflectometry (GNSS-R) for land applications is gaining a growing interest in recent years. The sensitivity of L-band, which is the GNSS operating frequency, to the water content of the observed targets has been largely proved: this suggests a potential of GNSS-R techniques for land applications. Many studies have been carried out by using ground based, airborne and satellite instruments for exploiting the GNSS-R ability for the retrieval of soil and vegetation parameters, such as the soil moisture (SM) and the aboveground biomass (AGB) (see e.g., Camps et al., 2016; Camps et al., 2020; Chew and Small, 2018; Clarizia et al., 2019; Carreno-Luengo et al., 2020; Guerriero et al., 2020; Santi et al., 2020). A comprehensive evaluation of the GNSS-R capabilities for land application is shown in Pierdicca et al., (2022), while the theoretical aspects are addressed in several other studies (see e.g. Dente et al., 2020). With respect to the microwave radiometers operating in the same frequency band, namely the SMAP and SMOS L-band radiometers that are routinely adopted for estimating SM and vegetation optical depth (VOD), GNSS-R has the further advantages of a relative independence of the variation in thermal

[*] Corresponding author.
E-mail address: e.santi@ifac.cnr.it (E. Santi).

https://doi.org/10.1016/j.srs.2024.100177
Received 13 September 2024; Received in revised form 4 November 2024; Accepted 15 November 2024
Available online 16 November 2024

background, which conversely affects the radiometric measurements. Thanks to the bistatic configuration, GNSS-R has also some advantages with respect to SAR: previous studies have shown GNSS-R signal saturation for AGB values of about 250–300 t/ha (Egido et al., 2014; Zribi et al., 2019), which is significantly higher than the 150 t/ha found for the L-band SAR (Dobson et al., 1992). On the other side, GNSS-R near specular reflections might exhibit low signal-to-noise (SNR) ratio in certain conditions (e.g., very dense vegetation and rough topography).

Focusing on soil moisture, Chew and Small (2018) proposed an algorithm based on multivariate regression to estimate SM from NASA's Cyclone GNSS (CyGNSS) (Ruf et al., 2015) data. Other implementations based on empirical models can be found in the literature: for instance, Yan et al. (2020) adopted a trilinear regression to address the effect of surface roughness in SM retrievals by using CyGNSS data, and Chew and Small (2018, 2020) related the SM variability to CyGNSS observables by using a linear regression. It is worth mentioning that the official CyGNSS SM product hosted by UCAR is based on such approach (Chew and Small, 2020). A different approach was proposed in Azemati et al. (2022), which used a bistatic forward model to retrieve SM from CyGNSS observations under various vegetation conditions, from bare soils to dense vegetation. Finally, Zribi et al. (2020) proposed a change detection algorithm which assumes incoherent return from the land surface.

Machine learning (ML) techniques have also been widely exploited to address the retrieval problem: despite the need of large datasets for training, these approaches showed very promising capabilities in obtaining accurate retrievals. Among others, the possibility of estimating SM by combining CyGNSS and ANN was exploited in (Eroglu et al., 2019), while Senyurek et al. (2020) based the SM retrieval on Random Forests (RF), and Nabi et al. (2022) exploited the full information contained in the DDM by using deep learning. Among the recent studies, Santi et al. (2022) attempted to map SM at global scale with a spatial resolution of 0.05° (≃5 km) by using CyGNSS data and ANN, while Hodges et al. (2024) proposed a blended soil moisture product obtained by combining five different CYGNSS soil moisture products based on a minimum variance estimator (MVE).

This growing interest on GNSS-R promoted the development of the HydroGNSS satellite mission, (Unwin et al., 2022), which has been selected as the second ESA Scout small satellite science mission, currently planned for launch in late 2024. HydroGNSS aims at operational mappings of SM, inundation or wetlands, freeze/thaw state and AGB, by also leveraging on the presence of a coherent channel, and on the capability of collecting reflections at both circular polarizations and from both GPS and Galileo constellations.

This study is devoted at assessing feasibility and limits of the ML based techniques for retrieving SM from GNSS-R, with the final aim of defining the base concepts for the HydroGNSS SM retrieval algorithm. In detail, two Machine Learning (ML) algorithms, namely ANN and RF, have been implemented, and their performances intercompared in terms of accuracy and computational cost.

In advance of HydroGNSS commissioning and operation, the analysis has been based on the NASA's CyGNSS observations, although the lack of double frequency and double polarization acquisitions, which will be available in HydroGNSS, do not allow quantifying the impact of this information on the retrieval accuracy.

Depending on the CyGNSS coverage, the target resolution for the SM product has been set to 0.05° (≃5 Km), which has been found as the upper limit of the trade-off with the revisit frequency. The resolution needs of course to be adapted to the different mission characteristics of HydroGNSS, which will be composed of two satellites operating in near polar orbit.

The main novelty of this study with respect to the existing literature is in the peculiar strategy that allows mapping SM at 0.05° resolution which, at the best of our knoweledge, is the highest resolution achieved for a CyGNSS derived SM product. The proposed approach assumes that the mechanism driving the scattering from vegetated soils is scale invariant, i.e. it does not change if moving from lower to higher resolution. This assumption allowed training the ML algorithms using as target the SMAP Level 3 SM daily global products (O'Neill et al., 2018) at 36 km resolution, and then applying the trained ML to CyGNSS data gridded at 0.05° to generate the SM output at this resolution. The latter has been finally validated against in-situ measurements from the International Soil Moisture Network (ISMN - Dorigo et al., 2011). The advantage of this approach is in the large availability of SM products at low resolution from several satellites (e.g. SMAP, SMOS, AMSR-2/3, ASCAT) that can be used as reference for training in case of unavailability of one or another sensor, while the SM availability at higher resolution is so far limited to the combined SMAP/Sentinel products (Das et al., 2020).

Moreover, the proposed implementation is independent of SMAP VOD/VWC and Roughness parameters that have been used as auxiliary inputs in most of the other studies involving CyGNSS and SMAP (e.g. Liu et al., 2023): this characteristic not only unlink the retrievals from SMAP data availability, coverage and spatial resolution, but especially avoids the conceptual weakness of using auxiliary input information (SMAP VOD and VWC) which is intrinsically correlated to the target SMAP SM.

Several combinations of CyGNSS observables and auxiliary data have been evaluated, including topography information, land use and vegetation biomass. This approach also allowed assessing the contribution of the CyGNSS observables other than the well assessed equivalent reflectivity (Clarizia et al., 2019) to the retrievals and quantifying the role of auxiliary information in improving the accuracy.

This study also analyses the accuracy dependence on land cover type, assess the predictor importance of each CyGNSS observable and auxiliary datum, and quantifies the relationship between training data amount and accuracy.

The paper is structured as follows: the description of test areas and datasets is provided in section II, the generation of the lower resolution (LR) and higher resolution (HR) datasets required by the method, is described in section III. The ML algorithm implementation is described in section IV. Sensitivity analysis and retrieval results at local and global scale are presented in section V and discussed in section VI.

2. Test areas and datasets

This study involved CyGNSS global data over land collected from August 2018 to July 2019, auxiliary information about topography, AGB and land use from various sources, and reference SM from SMAP L3 global daily products and ISMN.

A. Test areas and in-situ data (ISMN)

Data derived from ISMN within the same latitude range (±38°) and for the same time span of CyGNSS data have been considered with the twofold scope of analysing the CyGNSS sensitivity to SM and of validating the retrievals through comparison with in-situ measurements. The geographical distribution of the stations involved in the analysis is shown in Fig. 1. The hourly data from each station have been screened for quality (Dorigo et al., 2013) and then daily averaged and gridded on the same coordinates of CyGNSS and auxiliary data, to be comparable with the SMAP L3 SM and CyGNSS data.

A further validation at local scale has been also carried out by focusing on three networks in the ISMN database, namely SCAN in Walnut Gulch, Naqu, and OZNET, that have been also included in the SMAP core validation activities (Colliander et al., 2017).

Four stations of the SCAN network are available in ISMN for the Walnut Gulch watershed, in southeastern Arizona, U.S. The area is characterized by semi-arid conditions with brush and grass rangeland (Tolsdorf et al., 2021) and was already considered for cal/val activities of other satellite missions since the launch of AMSR-E.

The Naqu network is located in the Tibetan plateau (Su et al., 2011, 2013), which is characterized by cold climate, and it is mainly covered

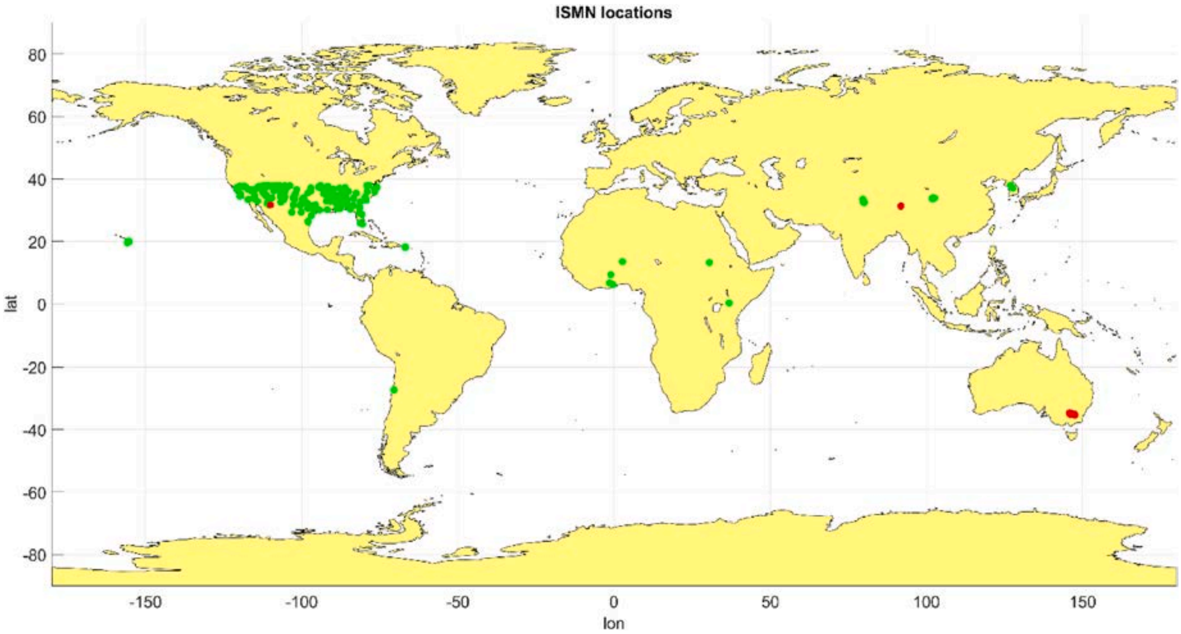


Fig. 1. Location of the ISMN stations involved in the validation. The position of all the stations is shown by green dots, except SCAN in Walnut Gulch, Naqu and OZNET networks that are shown in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

by grasslands. Seven stations from this network are available in ISMN.

The OZNET sites, in southeastern Australia (Panciera et al., 2014; Smith et al., 2012), are characterized by a climate between semiarid and temperate, with vegetation cover mainly composed by croplands and grasslands. Data from 18 OZNET stations from ISMN have been considered.

B. CyGNSS data

CyGNSS is a constellation of eight microsatellites operated by NASA and the University of Michigan (UM), that has been launched in December 2016. CYGNSS is observing the Earth surface in a latitude range between −38° and +38° with a sampling rate of 1 Hz (2 Hz since July 2019), by using a bi-static scatterometry technique based on Global Navigation Satellite Systems (GNSS) signals, thus making quasi-random the temporal and spatial resolution, since the first depends on the GNSS satellites in view and the second on the characteristics of the observed target (Zavorotny et al., 2014). The global dataset of CyGNSS Level 1b version v3.1 (the most recent available when this study was carried out) acquisitions over land collected between August 2018 and July 2019 has been downloaded from the dedicated data portal at Physical Oceanography Distributed Active Archive Center (PODAAC). Beside the reflectivity Γ, which has been computed according to Clarizia et al. (2019), the CyGNSS observables involved in the analysis are the SNR, the Effective Isotropic Radiated Power (EIRP), the normalized radar cross section (NRCS), the Kurtosis of the Delay Doppler Map (DDM) - as proxy of the dominant coherent/incoherent contribution - the peak power of the DDM (DDM PA), and the Trailing Edge width (TE) (Carreno-Luengo et al., 2020). In addition, the elevation angle served to filter out the acquisitions collected at low elevation (<45°) and as auxiliary input, to account for the dependence of received signals on observation geometry.

C. SMAP EASE-Grid SM

The SMAP L3 radiometer global daily 36 km EASE-Grid soil moisture V5 data have been downloaded from the National Snow and Ice Data Center (NSIDC) data portal for the same temporal period of the CyGNSS dataset.

Along with the globally gridded values of SM, provided as m^3/m^3, the vegetation water content (VWC – kg/m^2), and surface roughness (h - cm) products have been considered for understanding their effects on the GNSS-R observables. Data from ascending and descending overpasses in the same day have been averaged.

D. AGB pantropical map

The improved pan-tropical biomass map proposed by Avitabile et al. (2016) is used as auxiliary information to allow the algorithm accounting for the vegetation biomass effects in SM retrievals. The map contains AGB values in tons per hectare (t/ha), at 1 km resolution, in a latitude range between ±45°.

E. CCI Land Cover

The auxiliary information on land cover is derived from the ESA CCI Land Cover classification (LCC) in the latest version 2.0.7. The CCI LCC map is provided at 300 m spatial resolution as a classification of the Earth surface in 37 different classes, identified by a progressive number from 0 to 220 (https://www.esa-landcover-cci.org/?q=node/164).

F. GTOPO30 DEM

GTOPO30 is a global digital elevation model (DEM) provided with uniform 30 arc seconds grid spacing (approximately 1 km). The model is hosted by the U.S. Geological Survey (https://www.usgs.gov/media/files/gtopo30-readme.). Two parameters, namely the local elevation (DEM) and the local slope of the Earth surface (SLOPE), are considered here.

3. Dataset generation

The two LR and HR datasets required by the method have been generated by spatially and temporally co-registering all the available data and gridding them onto two common grids at 36 km (EASE-Grid 36) and 0.05° (≃ 5 km) spacing, respectively. Table 1 summarizes the CyGNSS and auxiliary data included in both datasets, the LR dataset has been further divided in training and test sets as described in section III.

Table 1
List of the CyGNSS and auxiliary data included in both datasets.

CyGNSS Observables:	Auxiliary data:
•EIRP	•Land Cover Classes (LCC) from ESA CCI
•Elevation Angle (EA)	•AGB from Pantropical map
•NRCS	•Surface elevation and slope (DEM + SLOPE)
•Kurtosis	•SM measurements from ISMN
•Γ	•SMAP L3 Radiometer global SM, VWC, *h (not in the 0.05° dataset)*
•SNR	
•Trailing Edge width	

A.LR dataset

The first dataset was obtained by coregistering all the data in the 36 km EASE-Grid reference system of the SMAP L3 product, with the scope of evaluating the CyGNSS sensitivity to SM, VWC, AGB and roughness parameter (h), as well as for training and testing the algorithms. The CyGNSS data have been preprocessed by filtering out low elevation (<45°) and two thresholds at 0.5 and 2 dB have been evaluated for SNR. Although SMAP retrievals in densely vegetated areas are not recommended (Entekhabi et al., 2014), SMAP data on dense forests like the Amazon and Congo rainforests have been maintained to extend the variability of observed conditions, which is a pivotal aspect for training the ML algorithms. The ISMN SM was also included by averaging the in-situ measurements from the stations within each EASE-Grid cell.

B. HR dataset

The second dataset was obtained by aggregating CyGNSS and auxiliary data on a grid at 0.05° (≃ 5 km) fixed spacing, with the scope of generating the output SM maps and validating them against ISMN. The same preprocessing of the EASE-Grid dataset has been applied to CyGNSS data, conversely, the HR dataset did not include the SMAP derived parameters that are not available at this resolution.

4. Retrieval algorithms

The SM mapping at 0.05° resolution using CyGNSS data has been based on the scale invariant assumption that the physical mechanism driving the scattering from vegetated soils does not change moving from lower to higher resolution. Given this assumption, the training, test and validation of the ML algorithms proposed in this study have been obtained through the following steps.

1) Training of the ML algorithms by using a subset of the LR dataset (36 km resolution).
2) Testing of the ML algorithms at 36 km on the remaining of the LR dataset not involved in training.
3) Applying the trained algorithms to the CyGNSS and auxiliary inputs from HR dataset to generate the SM output at 0.05° resolution.
4) Validating the SM at 0.05° resolution against ISMN.

Two popular ML techniques, namely ANN and RF, have been selected to implement the retrieval, and their results intercompared in terms of accuracy and computational cost with the aim of understanding which technique is more suitable for an operational application.

The ANNs have been largely applied to solve remote sensing problems (e.g., Dai et al., 2011; Del Frate et al., 2003; Elshorbagy and Parasuraman, 2008), thanks to their ability in approximating almost any kind of non-linear relationships (Hornik, 1989; Linden and Kinderman, 1989). The ANN implementation proposed in this study is based on the feed-forward multi-layer perceptron neural networks (MLP-ANN), with iterative optimization of hyperparameters based on Santi et al. (2016).

Like the ANNs, RF are gaining increasing popularity for solving remote sensing problems (e.g., Pal, 2005; Yu et al., 2018; Camargo et al., 2019; Marrs and Ni-Meister, 2019). RF belong to the ensemble learning methods (Breiman, 2001), which average the results coming from several weak predictors, also called decision trees, to establish the input/output relationship (Quinlan, 1993). An iterative optimization of the main hyperparameters, like the one adopted for ANN, has been carried out.

Both ANN and RF have been trained by considering CyGNSS observables and auxiliary data as inputs and SM as target. Ten different combinations of inputs have been implemented, with the aim of investigating the role of each CyGNSS observable and auxiliary parameter and assessing the contribution of observables other than ***Γ***: the iterative process including hyperparameters definition and training has been repeated for each combination. In detail, five input combinations included only ***Γ*** and SNR among the CyGNSS observables, and five included all the CyGNSS observables. To keep training and validation as much independent as possible, the algorithms have been tested at 36 km resolution against SMAP SM using the LR dataset and validated at 0.05° resolution against ISMN SM using the HR dataset.

Training and test sets have been obtained by systematically sampling the LR dataset: such sampling has been repeated three times by decreasing the amount of data for training from 20% of the total data (about 1.6 million data), to 10% (800.000 data), and to 1% (80.000 data). The algorithms have been trained over these data and tested on the remaining 6.4 million, 7.2 million and 7.9 million data, respectively. In terms of temporal coverage, with the 1% training - 99% test split, the algorithms have been trained over the equivalent of 4 days out of a year and tested on the remaining. An attempt to further decrease the training data amount down to 0.1% of the total dataset has been also carried out, but the poor results obtained suggested to consider 1% as the lower limit for the trade-off between accuracy and amount of training data.

The ANN “optimal” number of neurons and hidden layers and the transfer function type, between linear, tangent sigmoid (tansig) and logarithmic sigmoid (logsig), are defined according to the iterative search proposed by (Santi, 2016), which aims at preventing both overfitting and underfitting. Training of each configuration is repeated 20 times. The stop of each training run is ruled by the so-called Early Stopping (Prechelt, 1998) which also has the scope of preventing overfitting. Output of the optimization process were ANNs (one for each input configuration) composed by two hidden layers with the number of neurons increasing from a minimum of 6 neurons to a maximum of 24 neurons for each layer and a transfer function of type “logsig” or “tansig”, depending on the configuration. Expectably, the greater the number of inputs and the training data amount, the greater the number of neurons in the “optimal” configuration.

The training of RF was conducted in parallel to the one of ANN for each input configuration and using the same strategy. In this case, the hyperparameters that have been iteratively configured were the number of decision trees and the minimum number of leaf node observations (Marrs and Ni-Meister, 2019): the number of trees resulting from the optimization was 50, since lower values negatively affected the retrievals and higher values slowed down noticeably the training process without improving significantly the retrievals, and the minimum number of leaf node observations was 5. Among the other configurable hyperparameters in the Matlab ® implementation, the sampling with replacement has been enabled and the number of predictor variables for each decision split has been set to 4. The workflow of the hyperparameters definition and training for both ANN and RF is shown in Fig. 2.

After training and testing on the LR dataset, ANN and RF have been applied to the CyGNSS and auxiliary inputs from the HR dataset for generating the global SM maps at 0.05° resolution which have been validated against ISMN data. The top-level scheme of the overall implementation is shown in Fig. 3.

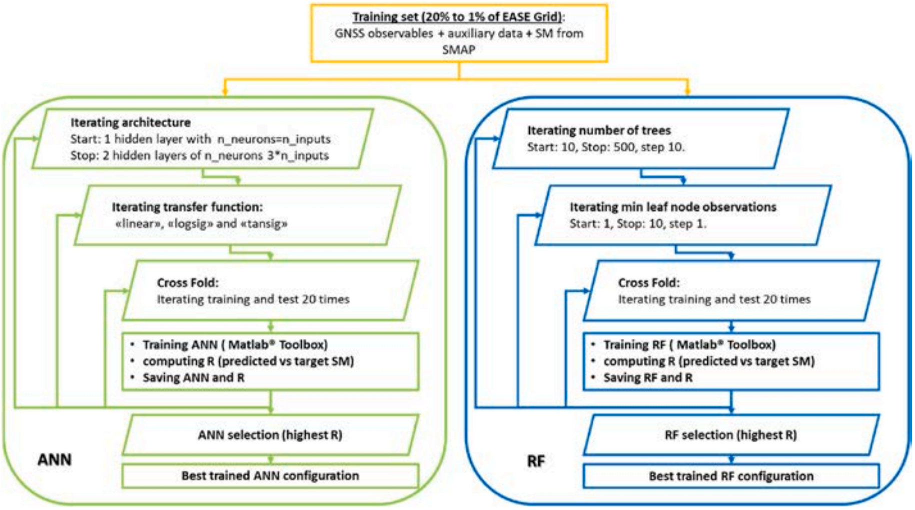

Training set (20% to 1% of EASE Grid):
GNSS observables + auxiliary data + SM from SMAP
Iterating architecture
Start: 1 hidden layer with n_neurons=n_inputs
Stop: 2 hidden layers of n_neurons 3*n_inputs
Iterating transfer function:
«linear», «logsig» and «tansig»
Cross Fold:
Iterating training and test 20 times
• Training ANN (Matlab® Toolbox)
• computing R (predicted vs target SM)
• Saving ANN and R
ANN selection (highest R)
Best trained ANN configuration
ANN
Iterating number of trees
Start: 10, Stop: 500, step 10.
Iterating min leaf node observations
Start: 1, Stop: 10, step 1.
Cross Fold:
Iterating training and test 20 times
• Training RF (Matlab® Toolbox)
• computing R (predicted vs target SM)
• Saving RF and R
RF selection (highest R)
Best trained RF configuration
RF

Fig. 2. The iterative hyperparameters definition and training of both ANN and RF.

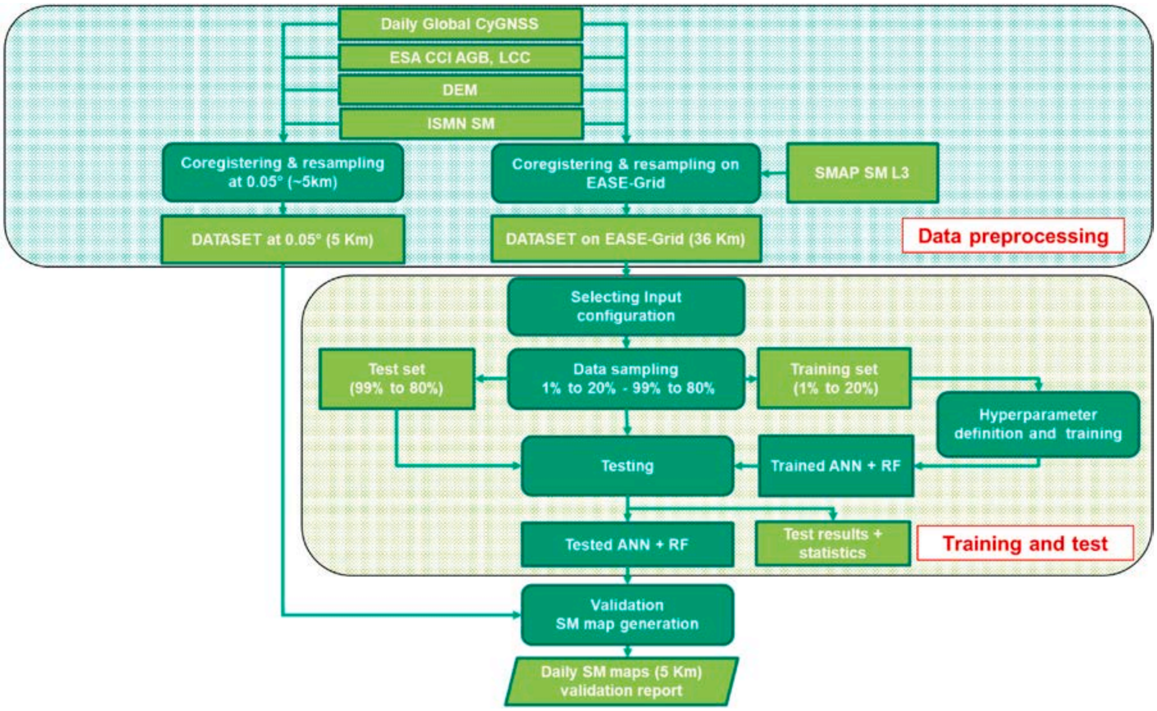

Daily Global CyGNSS
ESA CCI AGB, LCC
DEM
ISMN SM
Coregistering & resampling at 0.05° (~5km)
Coregistering & resampling on EASE-Grid
SMAP SM L3
DATASET at 0.05° (5 Km)
DATASET on EASE-Grid (36 Km)
Data preprocessing
Selecting Input configuration
Test set (99% to 80%)
Data sampling 1% to 20% - 99% to 80%
Training set (1% to 20%)
Hyperparameter definition and training
Testing
Trained ANN + RF
Tested ANN + RF
Test results + statistics
Training and test
Validation SM map generation
Daily SM maps (5 Km) validation report

Fig. 3. The SM retrieval algorithm flowchart, describing the training, test, and validation steps.

5. Results

A. Sensitivity Analysis

To better understand the relationship between CyGNSS observables and surface parameters, two sensitivity analyses have been carried out at 36 km and 0.05°, by comparing the CyGNSS observables to the surface parameters derived from SMAP and with the in-situ ISMN acquisitions, respectively. The analysis included the other parameters affecting the scattering mechanism, namely VWC, *h* and AGB.

5.1. Sensitivity analysis on EASE grid: CyGNSS vs. SMAP

The absolute values of correlation coefficients (R) between each of the CyGNSS observables and each of the target parameters are reported in Fig. 4: the tables are colour coded according to the R value from blue (lower R) to red (higher R).

The analysis has been carried out using the entire EASE-Grid dataset and the target parameters in the figures also include the roughness and VWC derived from SMAP, although these parameters have not been used as input for the retrieval algorithm since they are not available at a resolution suitable for being aggregated on the 0.05° grid. Fig. 4 a) refers to dataset filtered for SNR >0.5, while Fig. 4 b) refers to data filtered for SNR >2 dB, which is the most applied threshold value in literature (e.g. Clarizia et al., 2019).

Overall, the correlation between each CyGNSS observables and SM is insufficient for implementing a retrieval based on a single observable and linear relationships, thus supporting the use of ML techniques, which can exploit the synergy between multiple observables and leverage the nonlinear relationships between inputs and target for

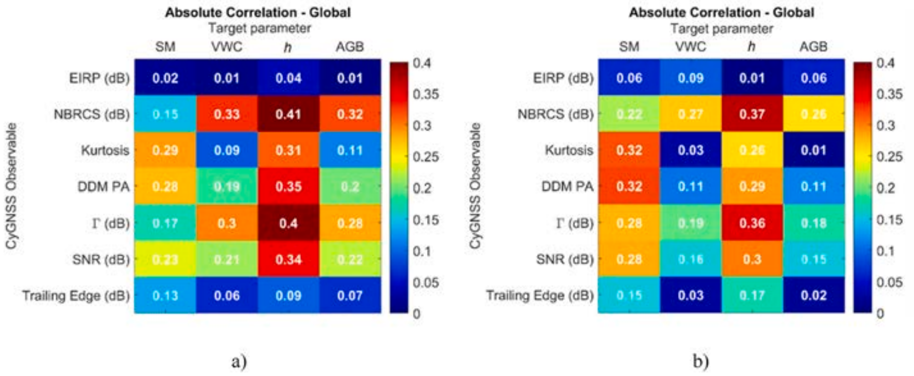


Fig. 4. Correlation between each of the CyGNSS observables, the SMAP derived SM, VWC, and Roughness, and the AGB from pantropical map: a) by thresholding data for SNR = 0.5 dB, b) by thresholding data for SNR = 2 dB.

improving the retrievals.

Among the CyGNSS observables, *Γ* and NBRCS show the highest correlation to biomass (either expressed by VWC or AGB), while some correlation to SM was pointed out by other observables as the Kurtosis and the DDM PA. However, test and validation of both ANN and RF implementations confirmed Γ as the most suitable parameter for the SM retrieval, while the other observables provided only negligible improvements.

The comparison between R values in Fig. 4 a) and 4 b) confirms that the higher the SNR threshold, the higher the sensitivity to SM. However, discarding the data with SNR $\leq$ 2 dB removed almost all data at higher AGB corresponding to forested areas.

Therefore, further effort has been spent in assessing the optimal values for SNR threshold, looking at the opposite needs of discarding data too much affected by noise and of keeping revisiting sufficient for operational retrievals.

5.2. *Sensitivity analysis at 0.05°: CyGNSS vs ISMN*

The sensitivity of CyGNSS observables has been assessed at global scale and 0.05° resolution against ISMN, with the twofold aim of confirming the findings at 36 km and of pointing out saturation effects (if any) that could set an upper limit to the retrievable SM.

The *Γ* sensitivity to SM is shown in Fig. 5: the scatterplot has been obtained by temporally averaging the hourly ISMN measurements collected within each day and by spatially averaging the network stations within each pixel at 0.05°, as specified in section III. To further assess the SNR thresholding criterion, the scatterplot is shown for two different threshold values: the blue dots in Fig. 5 represent the data with SNR >0.5 dB, while the red dots represent the subset of data with SNR >2 dB.

Correlation and sensitivity slightly improved with respect to those derived from the comparison with SMAP (Fig. 4). The scatterplot also points out a smaller data dispersion and an improved correlation to SM of the data with the higher SNR threshold (R = 0.4 vs. R = 0.31). However, the result has been obtained at the expense of coverage since the data amount decreased from about 22000 data points if using the 0.5 dB threshold down to about 16000 if using the 2 dB threshold. Finally, the analysis did not evidence any significant saturation effect in the relationship between *Γ* and SM, thus supporting the hypothesis that the SM retrieval can be achieved with similar accuracy on the entire range of SM, from 0 to ≃0.5 (m^3/m^3).

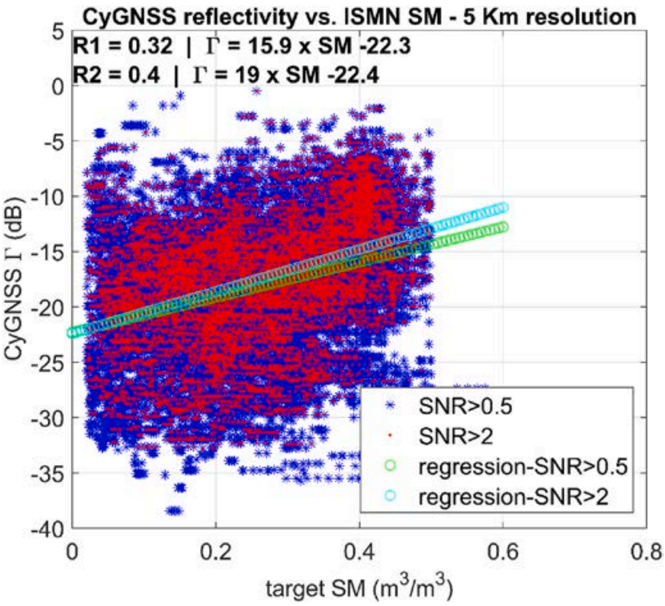


Fig. 5. CyGNSS Γ as a function of in-situ SM for the available dataset.

B. Algorithm test on EASE Grid

5.3. *SM algorithm test at global scale*

The R and RMSE values obtained when testing the ANN and RF on the subsets of EASE-grid dataset not used for training are summarized in Table 2 for the ten input combinations considered. Results refer to the algorithms trained using 1% of the total dataset and tested on the remaining 99%. The first five combinations only use Γ and SNR as CyGNSS inputs, the last five use all the CyGNSS observables derived from the DDM. The full list of inputs associated to each combination is provided in the first column of Table 2. These results have been obtained by setting the SNR threshold to 0.5 dB and including SNR in the inputs (Santi et al., 2022). This strategy did not cause any worsening of the retrieval accuracy with respect to the 2.0 dB threshold, with the advantage of significantly improving the coverage, especially in densely vegetated areas.

The obtained results demonstrate that the more auxiliary information is provided, the better is the result. On the other hand, the two different ML techniques are substantially equivalent since the ANN slightly outperforms RF in the configurations with less inputs and vice-versa RF outperforms ANN in the others.

Bias - not shown in Table 2 - was also very small, ranging between 10^{-4} m^3/m^3 at best and 10^{-3} m^3/m^3 at worst, with no clear prevalence of ANN or RF. This caused RMSE being practically overlapped to unbiased RMSE (ubRMSE), which is the other metric widely used for SM retrievals (Entekhabi et al., 2010), since the two quantities are related by the relationship: $ubRMSE^2 = RMSE^2\text{-}Bias^2$. For this reason, ubRMSE is not reported here.

Table 2
R and RMSE obtained by testing the ANN and RF algorithms against SMAP SM with the input combinations listed in the first column. Results refer to the algorithms trained on 1% of the total dataset and tested on the remaining 99%. The 5 upper lines of the table show the results obtained by using as CyGNSS input observables ***Γ*** and SNR only, the 5 lower lines show those obtained by using as inputs all the CyGNSS observables.

	ANN		RF	
	R	RMSE (m^3/m^3)	R	RMSE (m^3/m^3)
Γ + SNR + EA	0.44	0.130	0.42	0.135
Γ+SNR + EA + AGB	0.78	0.090	0.77	0.092
Γ+SNR + EA + AGB + Land cover	0.79	0.087	0.79	0.088
Γ+SNR + EA + AGB + Land cover + DEM + SLOPE	0.82	0.083	0.83	0.079
Γ+ SNR + EA + AGB + Land cover + DEM + SLOPE + lat + lon	0.86	0.073	0.88	0.068
Γ + SNR + Trailing Edge + Kurtosis + DDM PA + EA	0.54	0.120	0.53	0.122
Γ+SNR + Trailing Edge + Kurtosis + DDM PA + EA + AGB	0.80	0.087	0.79	0.087
Γ+SNR + Trailing Edge + Kurtosis + DDM PA + EA + AGB + Land cover	0.81	0.085	0.81	0.085
Γ+SNR + Trailing Edge + Kurtosis + DDM PA + EA + AGB + Land cover + DEM + SLOPE	0.83	0.081	0.84	0.078
Γ+SNR + Trailing Edge + Kurtosis + DDM PA + EA + AGB + Land cover + DEM + SLOPE + lat + lon	0.88	0.068	0.89	0.068

Conversely, the inclusion of CyGNSS observables other than Γ and SNR did not significantly improve the results, especially if the auxiliary information is fully exploited. In particular, the results obtained by the RF algorithm using the configuration with $\Gamma + SNR +$ all auxiliary inputs and the one with all CyGNSS observables + all auxiliary inputs configurations are practically overlapped, with R = 0.88 vs. R = 0.89, RMSE = 0.068 (m^3/m^3) and Bias = 0.00071 (m^3/m^3) for both.

As operational note, the different number of inputs has negligible effects on the computational cost in operational applications (i.e., application of the pre-trained algorithms to the datasets), while it affects the computational cost for training, since the configuration including all CyGNSS + all auxiliary data as inputs took about 75% more time for training than the corresponding configuration including only Γ and SNR among the CyGNSS observables + all auxiliary inputs.

If considering the other two combinations, 10% training and 90% test, and 20% training and 80% test, the accuracy improvement is small, as pointed out by the density plots of Fig. 6 a) and b) that show the results for the 20%–80% combination obtained by ANN and RF algorithms in the configuration with all CyGNSS + all auxiliary inputs.

In comparison with Table 2, R increases from 0.88 to 0.89 for ANN and from 0.89 to 0.91 for RF, while RMSE decreases from 0.068 m^3/m^3 to 0.066 m^3/m^3 for ANN and from 0.068 m^3/m^3 to 0.061 m^3/m^3 for RF. This suggests that the one percent of total dataset (≃80.000 input vectors which roughly correspond to 4 days over one year) is enough for the algorithms to learn the input/output relationship, while additional data do not bring information useful at improving the retrievals.

5.4. R, RMSE and relative error global maps

Statistics have been also computed for each grid cell in the EASE Grid dataset by comparing the timeseries of SM estimated by the ANN and RF algorithms with the SMAP reference. The results are shown in the maps of Fig. 7 for RF and Fig. 8 for ANN. Beside the already considered R and RMSE, the maps also show the mean of relative error (RE in %), computed as:

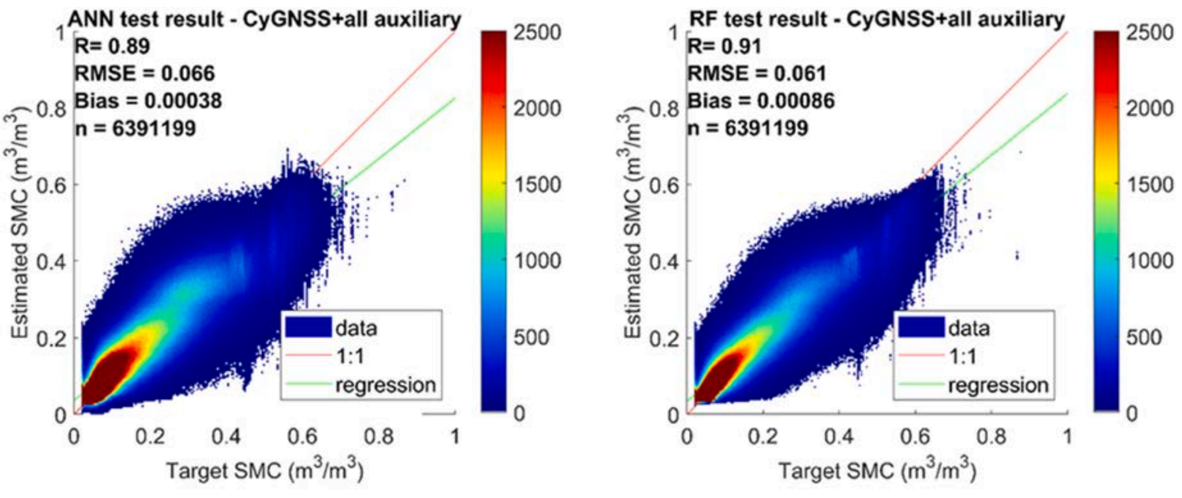
$$RE = \frac{SM_{CyGNSS} - SM_{SMAP}}{SM_{SMAP}} \quad (1)$$

Which has been introduced to further characterize the retrieval accuracy with respect to the SM dynamics.

The qualitative comparison of the results in Figs. 7 and 8 points out some better performance of RF in terms of R, that is less evident from the statistics in Table 2. Conversely, no outperforming retrieval method emerged if considering RMSE and RE.

To quantify the different performances, Fig. 9 a) shows the difference between RF and ANN RMSE maps of Figs. 7 b) and Fig. 8 b), while Fig. 9 b) shows the difference between the corresponding RE maps. The areas in which RF obtained lower RMSE than ANN are shown in blue, while the areas in which the ANN performed better than RF are shown in green.

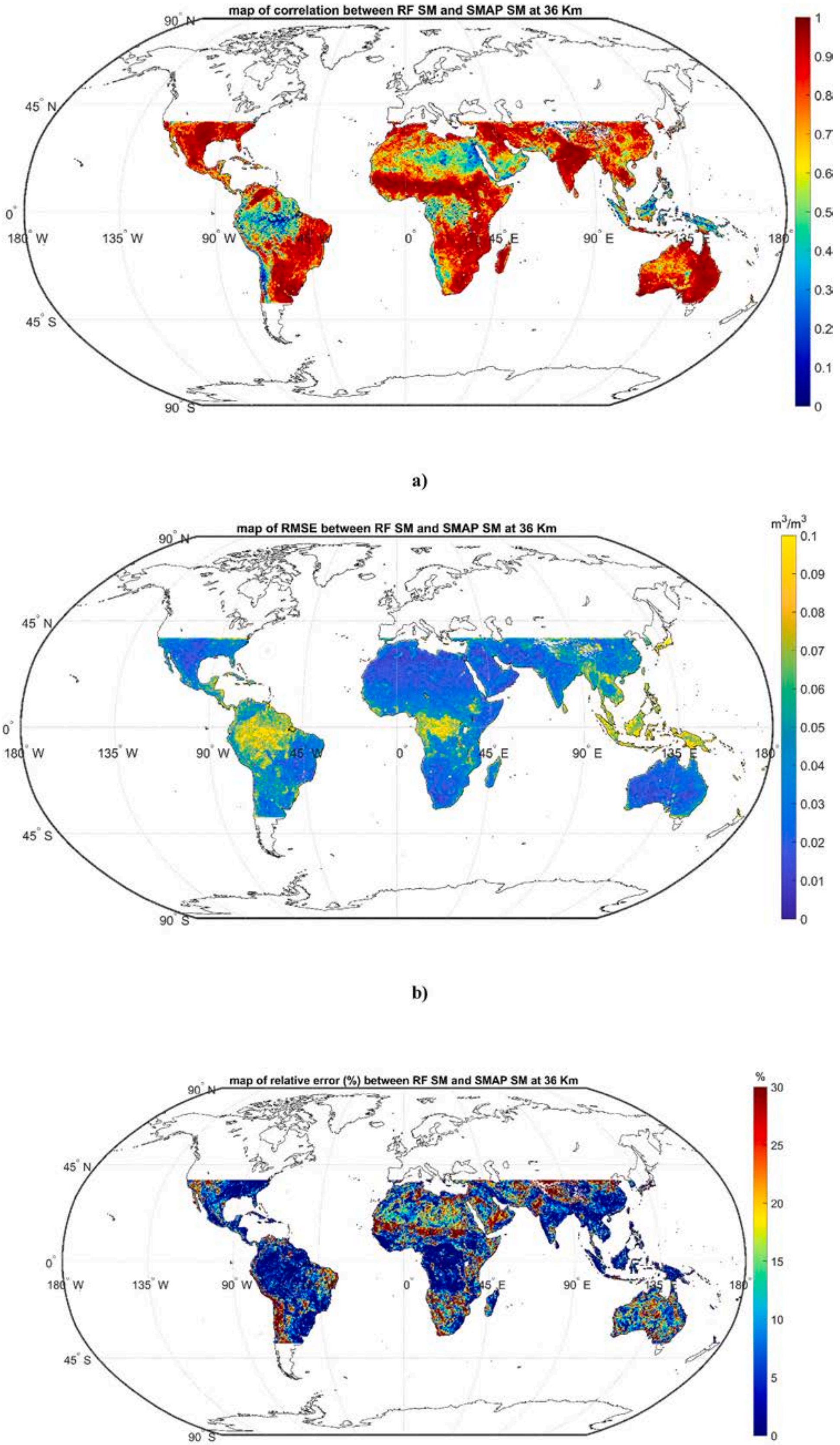
Overall, ANN has lower RMSE in about 40% of the pixels, mainly located in desert areas (e.g., northern, and southern Africa, central Australia) and equatorial forests (e.g., Central Africa, Amazon River basin), while RF obtains lower RMSE in the remaining. This result is reversed when looking at RE: by using this metric, ANN performs better than RF in 55% of the pixels, with no clear dependence on land cover: this also confirms how difficult is to establish a universal metric for evaluating the SM retrievals.



a)

b)

Fig. 6. Test result obtained with the 20%–80% split for a) ANN and all CyGNSS + all auxiliary, b) RF and all CyGNSS + all auxiliary input configurations.


map of correlation between RF SM and SMAP SM at 36 Km
90° N
45° N
0°
180° W
135° W
90° W
45° W
0°
45° E
90° E
135° E
180°
45° S
90° S
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
map of RMSE between RF SM and SMAP SM at 36 Km
m3/m3
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
map of relative error (%) between RF SM and SMAP SM at 36 Km
%
30
25
20
15
10
5
0

a)

b)

c)

Fig. 7. a) map of R computed over the entire timeseries between RF SM and SMAP SM at each grid cell of the EASE Grid dataset, b) the same for RMSE, c) the same for RE.

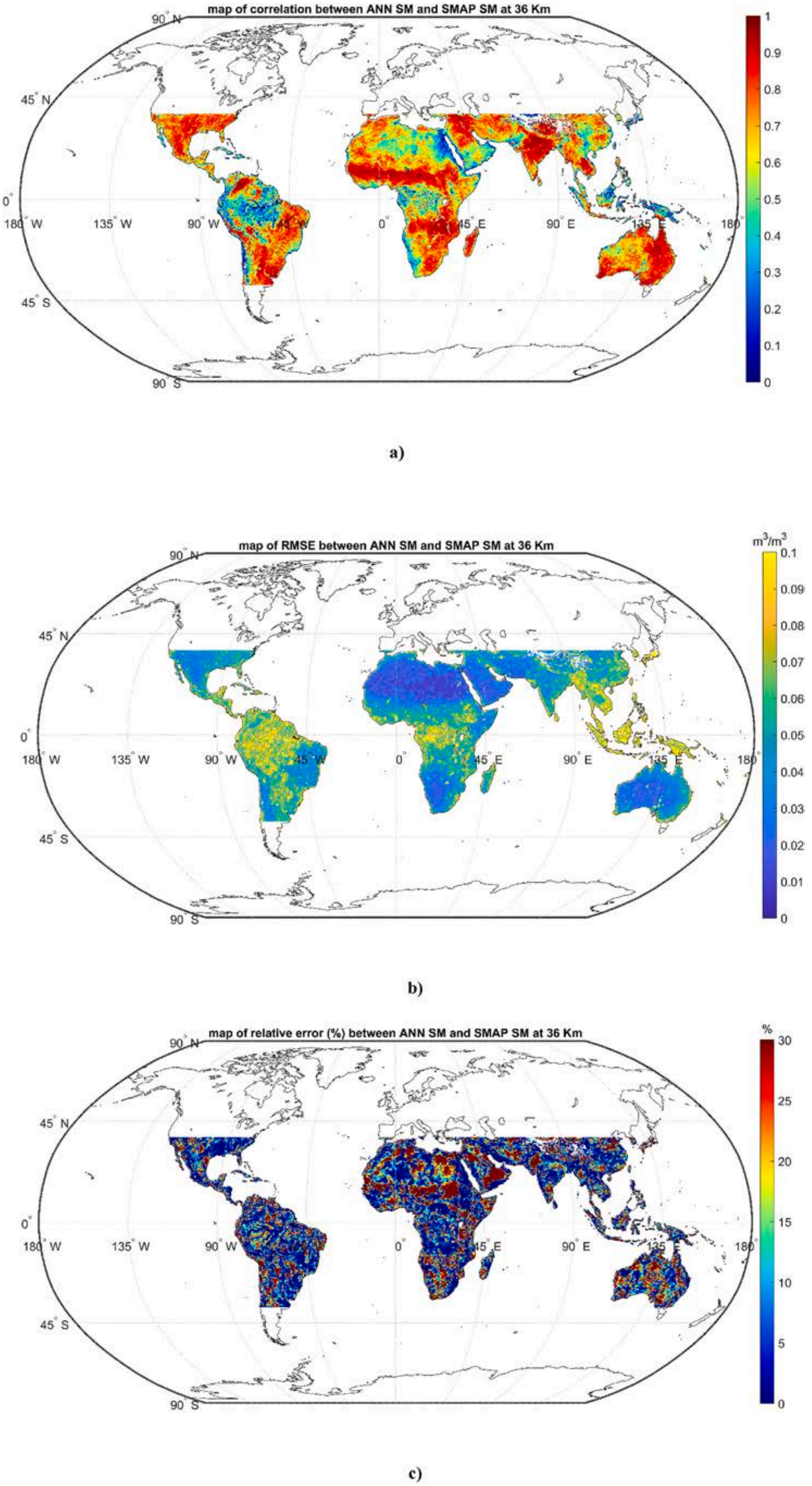

map of correlation between ANN SM and SMAP SM at 36 Km
90° N
45° N
0°
45° S
90° S
180° W
135° W
90° W
45° W
0°
45° E
90° E
135° E
180°
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
a)
map of RMSE between ANN SM and SMAP SM at 36 Km
m³/m³
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
b)
map of relative error (%) between ANN SM and SMAP SM at 36 Km
%
30
25
20
15
10
5
0
c)

Fig. 8. a) map of R computed over the entire timeseries between ANN SM and SMAP SM at each grid cell of the EASE Grid dataset, b) the same for RMSE, c) the same for RE.

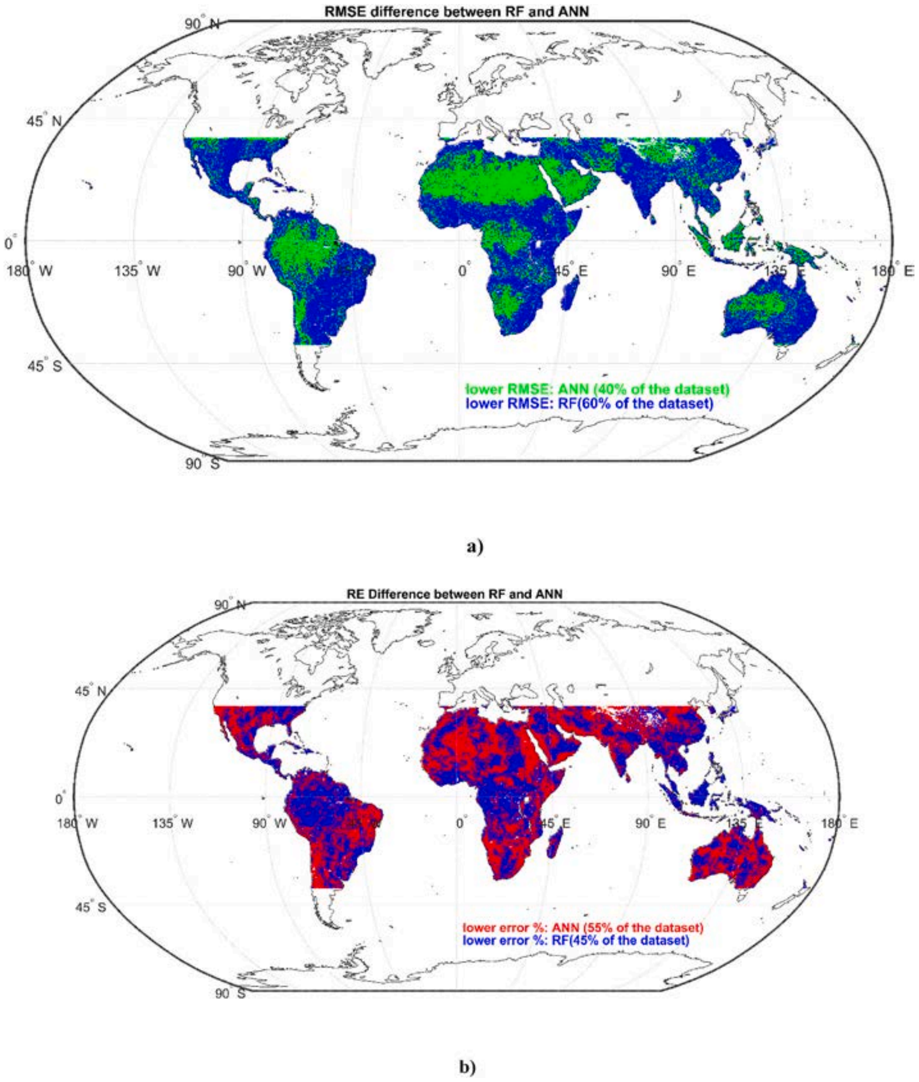

RMSE difference between RF and ANN
90° N
45° N
0°
180° W
135° W
90° W
0°
45° E
90° E
135° E
180° E
45° S
90° S
lower RMSE: ANN (40% of the dataset)
lower RMSE: RF(60% of the dataset)
a)
RE Difference between RF and ANN
90° N
45° N
0°
180° W
135° W
90° W
0°
45° E
90° E
135° E
180° E
45° S
90° S
lower error %: ANN (55% of the dataset)
lower error %: RF(45% of the dataset)
b)

Fig. 9. Comparison between RF and ANN results a) in terms of RMSE and b) RE.

Figs. 7 and 8 also point out the accuracy dependence on the geographical area and land cover. The highest R and lowest RMSE are obtained for the areas characterized by scarce to low vegetation (apart from deserts), while the disagreement between CyGNSS derived and SMAP derived SM is more pronounced in the equatorial forests of south America and Africa. The accuracy also decreases when getting close to the ±38° latitude that represents the limit of the CyGNSS coverage, because of the decrease of valid acquisitions due to SNR and elevation thresholding.

5.5. Dependence on land cover

The results shown in Figs. 7 and 8 suggested to further characterize the retrieval accuracy as a function of land cover. For this analysis, the dataset has been divided based on the CCI land cover map.

The original 37 classes in the CCI map have been aggregated in 6 classes, including deciduous and evergreen forests, bare and scarcely vegetated surfaces, cultivated fields, natural short vegetation, and mixed forest/short vegetation, plus a class including the remaining. The results obtained by ANN and RF for each of the 6 classes are summarized as R and RMSE values in Table 3.

Bias was between 10^{-4} and 10^{-3} m^3/m^3 for all the classes. These results have been obtained by using the input configuration including all CYGNSS observables + all auxiliary data.

Table 3
Test results for ANN and RF grouped by different land cover classes.

ANN results				
	R	RMSE (m^3/m^3)	RE (%)	data #
Cultivated	0.8	0.073	17.7	1199237
Mixed forest/short vegetation	0.81	0.086	19.3	318431
Forest evergreen	0.67	0.100	9.1	862619
Forest deciduous	0.71	0.080	16.9	488609
Natural short vegetation	0.83	0.058	16.3	1678834
Bare soils/scarce vegetation	0.71	0.037	20.7	2261229
RF results				
	R	RMSE (m^3/m^3)	RE (%)	data #
Cultivated	0.79	0.075	24.4	1199237
Mixed forest/short vegetation	0.80	0.088	24.2	318431
Forest evergreen	0.67	0.101	8.3	862619
Forest deciduous	0.68	0.084	22.2	488609
Natural short vegetation	0.83	0.060	21.2	1678834
Bare soils/scarce vegetation	0.75	0.036	19.8	2261229

As expected, the lowest RMSE is obtained for the bare/scarcely vegetated soils. Natural vegetation and cultivated fields obtain the highest R and quite low RMSE, while the accuracy worsened on both evergreen and deciduous forests, for which L-band is evidently not sufficient and lower frequencies would be necessary. Notably, the forest evergreen class, which was characterized by the highest RMSE, also showed the lowest RE, because of the high SM range (SM mean ≃ 0.35 m^3/m^3). For the other classes, which are all characterized by a mean SM below 0.2 m^3/m^3, the behaviour of RE roughly agrees with the one of RMSE.

5.6. Dependence on forest biomass

The analysis by land cover pointed out some accuracy decrease in forested areas. This can be attributed to the "disturbing" effect of vegetation biomass on the measured signals that is not completely compensated by the use as auxiliary input of the information from AGB map. This aspect has been further analysed for better understanding the GNSS-R limits in mapping SM in forested areas. To this aim, the relationship between R and RMSE from Figs. 7 and 8 and the AGB data has been studied, finding some decreasing behaviour of R and increasing of RMSE when AGB increases. As an example, the RMSE obtained by the ANN algorithm (see Fig. 8) is plotted as a function of AGB in Fig. 10: the RMSE increase with AGB is evident from the plot, although the data are quite dispersed (R = 0.35). RF obtained similar behaviours and correlations (not shown).

In evaluating these results, it should be however noticed that the SMAP SM product considered as target for computing the RMSE could be similarly affected by vegetation, since also SMAP operates in the same frequency band, thus biasing the results.

5.7. Contribution of CyGNSS and auxiliary data

The relative contribution of CyGNSS observables and auxiliary inputs to the overall result has been quantified for both ML implementations by focusing on the "CyGNSS + all auxiliary" configuration (Table 2). The impact of each predictor on the results has been computed for RF according to Breiman (2001), by verifying the retrieval accuracy with algorithm reruns using only subsets of inputs randomly combined. For ANN, the predictor importance has been obtained by deriving the total weight applied to each of the input parameters from the combination of weights applied during its propagation in the ANN.

Fig. 11 shows the relative (%) contribution of CyGNSS observables and auxiliary data to the SM estimation for RF (Fig. 11a) and ANN (Fig. 11 b).

The results in Fig. 11 demonstrate that in both ANN and RF implementations, the SM retrieval is driven by CyGNSS observables. The scope of including the auxiliary data is in constraining the inversion, by helping the ML algorithms in discriminating between different targets having the same electromagnetic response. The small differences (a few %) in the weight of the auxiliary data between the two regressors can be attributed to the peculiar ANN and RF architectures and training rules which are based on different statistical concepts that consequently explore the space of solutions in different ways.

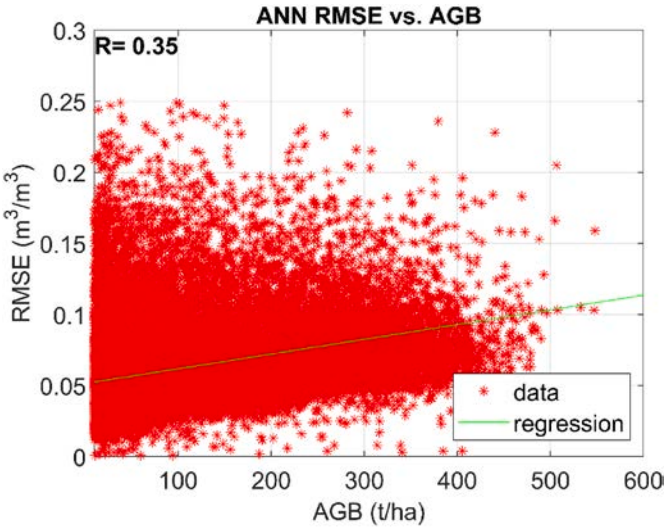


Fig. 10. RMSE global values from the map in Fig. 8 as a function of AGB.

C. Algorithm validation at 0.05°

The validation of both ANN and RF has been carried out at 0.05° against ISMN data at global scale, by considering the entire ISMN dataset available, and at local scale, by focusing on the three sites described in section II that have been already considered as core validation sites for other satellite missions.

5.8. SM at 0.05° against ISMN: global

The SM algorithm validation has been carried out by comparing the SM estimated by the ANN and RF algorithms with the in-situ SM obtained by averaging the ISMN measurements available within each grid cell at 0.05° resolution. The results in terms of R, RMSE and Bias are summarized in Table 4. As a term of comparison, the statistics obtained by comparing the SMAP SM with the ISMN data aggregated on EASE Grid are also reported in Table 4.

Although in line with the sensitivity shown in Fig. 5, the results listed in Table 4 seem pointing out some poor retrieval performances. However, also SMAP results are far from the official statistics for L3 product, especially for RMSE. This suggests the presence in the dataset of ISMN networks of stations with anomalies not identified in the pre-screening, or simply not representative for the given pixel size. A one-by-one control of the 6100 datasets contained in the ISMN (Dorigo et al., 2013) is not feasible, suggesting to focus on selected test sites that have already been considered for validation of other satellite products, as those described in section II. Anyway, the slightly better results obtained by ANN and RF at 0.05° with respect to SMAP at 36 km seem suggesting some potential of the proposed retrieval technique.

5.9. SM at 0.05° against ISMN: local

The validation results for the selected ISMN networks of Walnut Gulch, NAQU and OZNET are summarized in Table 5 and shown in Fig. 12. These results have been obtained by comparing the SM output of ANN and RF with the target SM obtained by aggregating the ISMN stations within each pixel at 0.05°. The results obtained by comparing SMAP SM with ISMN aggregated on EASE Grid are also reported for comparison. Validation only included the dates in which both CyGNSS and SMAP acquisitions were available.

The different spatial resolution prevents the direct comparison between ANN/RF and SMAP. Anyway, the statistics of Table 5 show some improvement when moving from 36 km to 0.05°. This is also evident from the scatterplots of Fig. 12, which show the retrieved SM as a function of the target SM. The results obtained at 0.05° by ANN are displayed in the first column, those obtained by RF in the second column, while SMAP is shown the third column. In Walnut Gulch, RF outperforms both ANN and SMAP, although some overestimation of low and underestimation of high SM, appears, causing the regression line to be farther from the 1:1. Notably, RMSE is in all cases within the 0.04 m^3/m^3 target accuracy for satellite products.

In MAQU, SMAP is very well correlated to ISMN (R = 0.91), although the RMSE is also very high (>0.11 m^3/m^3), because of evident overestimation of the high SM. Such overestimation is not shown by ANN and RF that exhibit an appreciable decrease of RMSE. Both methods show however an anomalous behaviour for the low SM values. In detail, RF shows some saturation and ANN a large dispersion of the data that could be attributed to residual presence of frozen soil which was not completely filtered out during the data preprocessing. In this case,

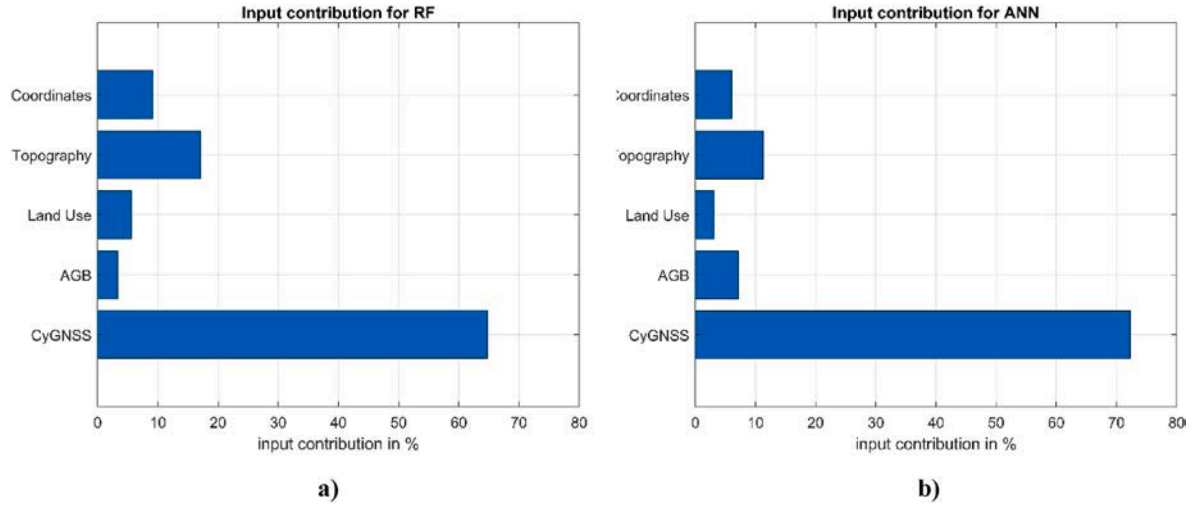


Fig. 11. Total contribution of CyGNSS and auxiliary data for the SM retrieval a) using RF and b) using ANN.

Table 4
Main statistics of the global validation against ISMN.

	ANN (0.05°)	RF (0.05°)	SMAP (EASE 36)
R	0.64	0.65	0.57
RMSE (m^3/m^3)	0.082	0.084	0.087
Bias (m^3/m^3)	0.005	0.013	0.012

Table 5
Main statistics of the validation against selected ISMN stations.

	R			RMSE (m^3/m^3)		
	ANN	RF	SMAP	ANN	RF	SMAP
Walnut Gulch	0.60	0.78	0.74	0.037	0.029	0.034
MAQU	0.71	0.79	0.91	0.051	0.057	0.114
OZNET	0.60	0.70	0.67	0.068	0.065	0.059

neither ANN nor RF clearly outperforms each other.

In OZNET, RF obtains a slightly better correlation than SMAP (R = 0.7 vs. R = 0.67) while RMSE is substantially equivalent. Worse results are obtained by ANN (R = 0.6 and RMSE = 0.07). Both RF and ANN regressions are closer to the 1:1 line, while SMAP tends to overestimate the higher SM and exhibits higher data dispersion than RF.

D. SM maps: CyGNSS 0.05° vs. SMAP 36 km

After test and validation, the ANN and RF algorithms have been applied to generate daily global SM maps at 0.05° resolution from the available data. Given the quasi-random coverage, the maps have been generated by considering the most recent CyGNSS acquisitions in a 3-day temporal window for each pixel in the grid. This approach has been preferred to the spatial interpolation (e.g. by Chew, 2021) to better understand the revisiting. Besides the already mentioned validation against in-situ SM, a quantitative validation of the maps is difficult, since reference SM data are almost impossible to find at this resolution and global scale, except for the merged SMAP/Sentinel SM products (Das et al., 2020) that however could not be completely independent of the SMAP 36 km data used for training, thus affecting the results.

To evaluate the resolution improvement, a qualitative comparison with the original SMAP SM at 36 km has been attempted. Examples are shown in the maps of Fig. 13: Fig. 13 a), left panel, shows a SM map derived from CyGNSS and RF at 0.05° resolution for an area of 10° × 10° that includes the Walnut Gulch watershed in USA. The result has been obtained for November 8th, 2018; the location of the SCAN stations considered for validating the algorithm is shown in the map as red dots. Fig. 13 a), right panel, shows the corresponding L3 SMAP SM.

Fig. 13 b) left panel shows instead another area of 10° × 10° that includes the OZNET network for May 27th, 2019: on the left the CyGNSS derived SM and on the right the corresponding SMAP data. Finally, Fig. 13 c) left shows the CyGNSS derived SM at 0.05° for the Tibetan plateau obtained in date May 28th, 2019, while Fig. 13 c) right shows the corresponding SMAP SM at 36 km.

In all cases, the SM maps at 0.05° exhibit slightly higher SM values than those reported by SMAP in the driest areas. Besides this, a qualitative agreement between the SM patterns identified by SMAP at lower resolution and those identified by CyGNSS at higher resolution appears from the comparison; moreover, CyGNSS at 0.05° seems capable to provide finer details and to identify additional patterns at small scale. Unfortunately, the unavailability of SM distributed data at 0.05° resolution to be used as reference, prevented any attempt to further quantify this comparison.

6. Discussion

The results shown in section V offered several ideas and directions for the development of the Level 2 SM algorithms for the ESA's Scout-2 HydroGNSS mission.

Test and validation confirmed the effectiveness of the proposed technique, that makes use of reference SM at lower resolution to generate global SM maps at higher resolution, thus overcoming the lack of distributed SM data at 0.05° to be used as the reference for training. In this sense, the proposed workflow (Fig. 3) can be easily adapted to HydroGNSS: the ML inputs can be easily modified, added or removed and re-training on the new datasets is not an issue considering the computational resources of recent machines.

Concerning the observables to be included in the retrieval, the reflectivity has been confirmed as the main driving observable, while negligible improvements have been obtained if adding the other observables except for SNR. Adding the latter observable to the ML inputs allows relaxing the SNR threshold from the conventionally adopted 2 dB down to 0.5 dB, with negligible effects on the accuracy and sensible advantages for the coverage.

Concerning the accuracy, the results did not point out any relevant saturation for the higher SM values, confirming that the entire SM range can be retrieved with similar accuracy. In this sense a prevailing technique between ANN and RF did not emerge from the analysis. Moreover, the tests with different input combinations also pointed out that the auxiliary “static” information, like topography or land cover, has a not negligible role in bounding the inverse problem. This solution appears

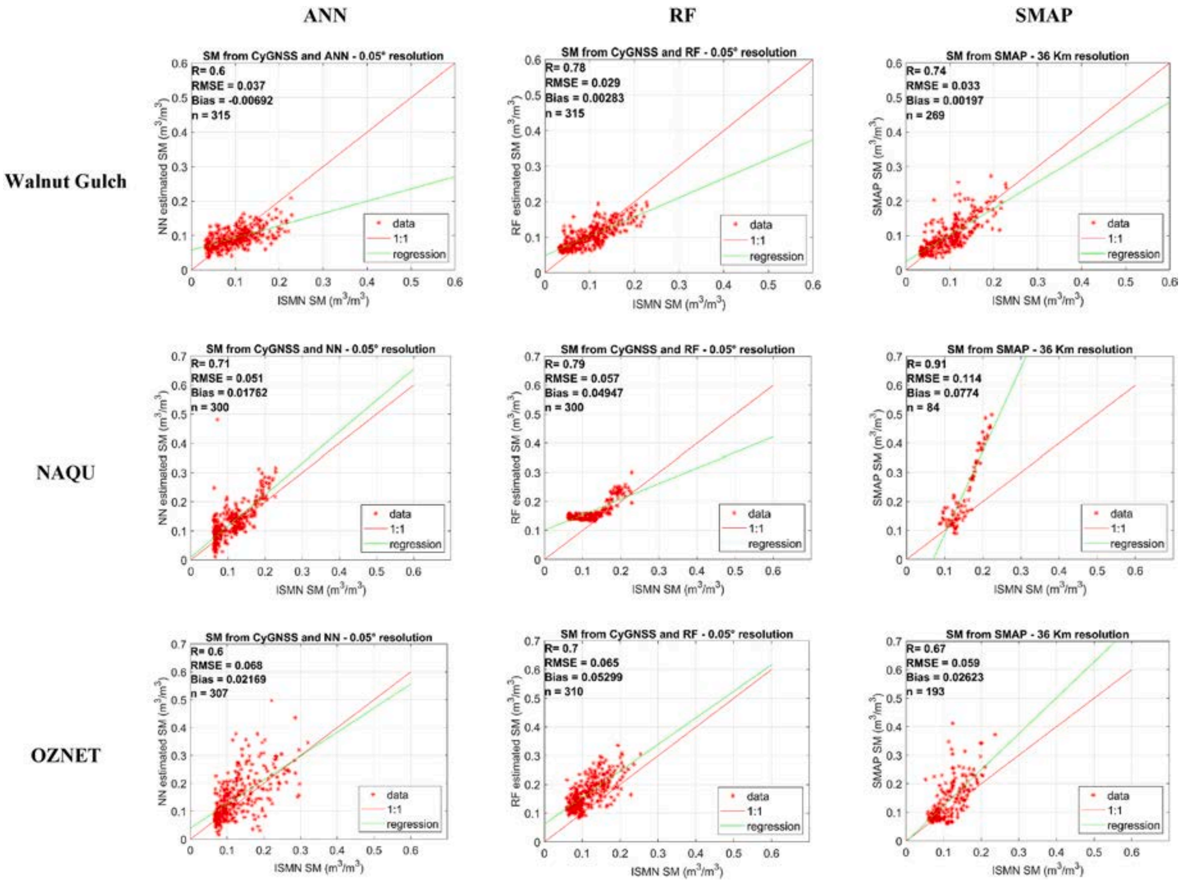


Fig. 12. Validation at 0.05° in the three ISMN sites obtained by ANN (1st column) and RF (2nd column). SMAP SM is shown in the 3rd column.

better suitable than other proposals involving auxiliary "dynamic" information also derived from SMAP, as e.g. VWC or VOD, with the disadvantages of linking resolution and coverage to those of SMAP, and making the retrievals biased by the intrinsic correlation (up to $R = 0.75$ in the dataset considered for this study) between SM and the VWC/VOD that are all part of the same processing.

Related to the vegetation effect, the analysis pointed out some accuracy decrease in forested areas because of the attenuation effect of vegetation biomass on the CyGNSS observation that the use of the AGB map as auxiliary input is not sufficient to compensate completely. This aspect deserves further analysis, which will be carried out in the prosecution of this research for better understanding the GNSS-R limits in mapping SM in densely forested areas as e.g. the Congo and Amazon rainforests.

Focusing on the results that can be transferred to HydroGNSS, this study cannot provide final answers on the maximum spatial resolution achievable, being HydroGNSS planned to operate with two satellites in near polar orbit, thus with completely different mission characteristics.

Also, this study cannot quantify the contribution of dual frequency and dual polarization data, that are peculiar of HydroGNSS, to the retrievals: this analysis is so far demanded to model simulations (Dente et al., 2020, 2024). Further analysis shall also address the aspects related to the higher latitudes that HydroGNSS will cover, as the freeze/thaw and its effects on the SM retrievals in boreal areas. To date, these aspects have been partially investigated in Santi et al. (2020) based on TechDemoSat-1 (TDS-1 –(Unwin et al., 2016)).

7. Conclusions

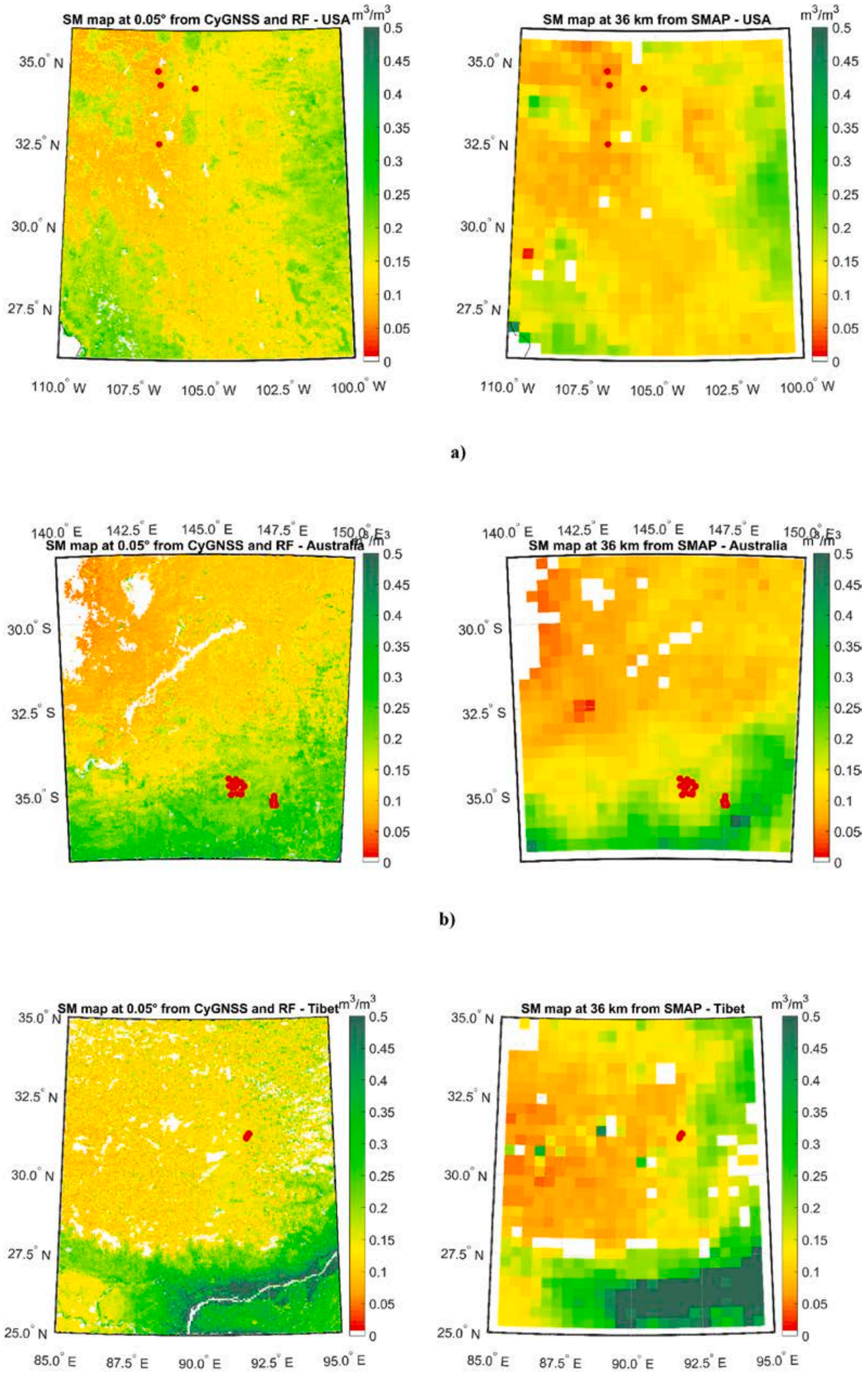
Two machine learning techniques aimed at exploiting GNSS-R for the retrieval of Soil Moisture have been developed and validated by using CyGNSS data, with the aim of defining retrieval concepts to be applied to the incoming HydroGNSS mission.

Based on the obtained results, GNSS-R is confirmed as a suitable and promising tool for mapping SM at global scale and few kilometres resolution. Considering that both satellite missions aiming at this scope, namely SMAP and SMOS, are well beyond their lifetime, and no heir missions with comparable resolution are foreseen so far, GNSS-R could represent the sole opportunity for global SM mapping in a near future. In this sense, it should be remarked that the proposed implementation makes the operational application of the trained algorithms independent of the availability of SMAP (or any other EO) SM product, which is only needed for training.

Further efforts should be carried out to verify and improve the generalization capabilities of the proposed algorithms and to identify alternative sources of reference SM to update the training in case the abovementioned missions will cease operations. The possibility of lowering the 0.06 m^3/m^3 RMSE, which appears so far as the accuracy limit for both CyGNSS and SMAP in global scale retrievals, will also be evaluated in the prosecution of this research. Further analysis at regional and global scale will be carried out in this sense for better understanding the peculiarities of the bistatic scattering mechanism and addressing the effect of confounding factors as roughness, vegetation and inland waters, with the final aim of finding accuracy improvement strategies that can be transferred to HydroGNSS.

CRediT authorship contribution statement

Emanuele Santi: Writing – original draft, Validation, Software, Methodology, Formal analysis, Conceptualization. **Davide Comite:** Writing – review & editing, Data curation. **Laura Dente:** Writing – review & editing, Data curation. **Leila Guerriero:** Writing – review &



(caption on next page)

Fig. 13. Sample global maps at 0.05° resolution generated by the RF algorithm (left column) compared with the corresponding SMAP SM product at 36 km resolution (right column): a) USA – Walnut Gulch for November 5th, 2018, b) Australia – OZNET for May 27th, 2019, c) Tibetan Plateau -NAQU for May 28th, 2019. The red markers correspond to the location of in-situ ISMN stations. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

editing, Supervision, Conceptualization. **Nazzareno Pierdicca:** Writing – review & editing, Supervision, Data curation, Conceptualization. **Maria Paola Clarizia:** Writing – review & editing, Data curation, Conceptualization. **Nicolas Floury:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study has been funded by the European Space Agency (ESA) under the contracts Exploiting CYGNSS GNSS-R over Land and Ocean for GeophYsics (ECOLOGY) – ESA/ESTEC CONTRACT n. 4000130560/20/NL/FF/gp/ and Scout-2 programme HydroGNSS - ESA Subcontract No: MS000434.

Data availability

Data will be made available on request.

References

Avitabile, V., Herold, M., Heuvelink, G., Lewis, S.L., Phillips, O.L., Asner, G.P., et al., 2016. An integrated pan-tropical biomass maps using multiple reference datasets. Global Change Biol. 22, 1406–1420. https://doi.org/10.1111/gcb.13139.

Azemati, A., Melebari, A., Campbell, J.D., Walker, J.P., Moghaddam, M., 2022. GNSS-R soil moisture retrieval for flat vegetated surfaces using a physics-based bistatic scattering model and hybrid global/local optimization. Rem. Sens. 14, 3129. https://doi.org/10.3390/rs14133129.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Camargo, F.F., Sano, E.E., Almeida, C.M., Mura, J.C., Almeida, T., 2019. A comparative assessment of machine-learning techniques for land use and land cover classification of the Brazilian tropical savanna using ALOS-2/PALSAR-2 polarimetric images. Rem. Sens. 11, 1600.

Camps, A., Park, H., Castellvi, J., Corbera, J., Ascaso, E., 2020. Single-pass soil moisture retrievals using GNSS-R: lessons learned. Rem. Sens. 12, 2064.

Camps, A., Park, H., Pablos, M., Foti, G., Gommenginger, C., Liu, P.-W., Judge, J., 2016. Sensitivity of GNSS-R spaceborne observations to soil moisture and vegetation. IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. 9 (10), 4730–4741.

Carreno-Luengo, H., Luzi, G., Crosetto, M., 2020. Above-ground biomass retrieval over tropical forests: a novel GNSS-R approach with CyGNSS. Rem. Sens. 12. https://doi.org/10.3390/rs12091368.

Chew, C., Small, E., 2020. Description of the UCAR/CU soil moisture product. Rem. Sens. 12, 1558.

Chew, C., Small, E., 2018. Soil moisture sensing using spaceborne GNSS reflections: comparison of CYGNSS reflectivity to SMAP soil moisture. Geophys. Res. Lett. 45, 4049–4057.

Chew, C., 2021. Spatial interpolation based on previously observed behavior: a framework for interpolating spaceborne GNSS-R data from CYGNSS. Spatial Sci. 68 (1), 155–168. https://doi.org/10.1080/14498596.2021.1942253.

Clarizia, M.P., Pierdicca, N., Costantini, F., Floury, N., 2019. Analysis of CYGNSS data for soil moisture retrieval. IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. 12 (7), 1–9. https://doi.org/10.1109/JSTARS.2019.2895510.

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

Dai, X., Huo, Z., Wang, H., 2011. Simulation for response of crop yield to soil moisture and salinity with artificial neural network. Field Crops Res. 121 (3), 441–449. https://doi.org/10.1016/j.fcr.2011.01.016.

Das, N., Entekhabi, D., Dunbar, R.S., Kim, S., Yueh, S., Colliander, A., O'Neill, P.E., Jackson, T., Jagdhuber, T., Chen, F., Crow, W.T., Walker, J., Berg, A., Bosch, D., Caldwell, T., Cosh, M., 2020. SMAP/Sentinel-1 L2 radiometer/radar 30-second scene 3 km EASE-grid soil moisture. NASA National Snow and Ice Data Center Distributed Active Archive Center. https://doi.org/10.5067/ASB0EQO2LYJV [Data Set]. Boulder, Colorado USA, Version 3.

Del Frate, F., Ferrazzoli, P., Schiavon, G., 2003. Retrieving soil moisture and agricultural variables by microwave radiometry using neural networks. Remote Sens. Environ. 84 (2), 174–183. https://doi.org/10.1016/S0034-4257(02)00105-0.

Dente, L., Guerriero, L., Comite, D., Pierdicca, N., 2020. Spaceborne GNSS-R signal over a complex topography: modelling and simulations. IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. 13 (1), 1218–1233.

Dente, L., Comite, D., Guerriero, L., Zribi, M., Pierdicca, N., 2024. Polarimetric features of GNSS-R signal over land: a simulation study. IEEE Trans. Geosci. Rem. Sens. https://doi.org/10.1109/TGRS.2024.3409880. Early Access.

Dente, L., Guerriero, L., Comite, D., Pierdicca, N., 2020. Space-borne GNSS-R signal over a complex topography: modeling and validation. IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. 13, 1218–1233.

Dobson, M.C., Ulaby, F.T., LeToan, T., Beaudoin, A., Kasischke, E.S., Christensen, N., 1992. Dependence of radar backscatter on coniferous forest biomass. IEEE Trans. Geosci. Rem. Sens. 30 (2).

Dorigo, W.A., Xaver, A., Vreugdenhil, M., Gruber, A., Hegyiová, A., Sanchis-Dufau, A.D., Zamojski, D., Cordes, C., Wagner, W., Drusch, M., 2013. Global automated quality control of in situ soil moisture data from the international soil moisture network. Vadose Zone J. 12 (3). https://doi.org/10.2136/vzj2012.0097 vzj2012.0097.

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

Egido, A., Paloscia, S., Guerriero, L., Pierdicca, N., Motte, E., Santi, E., Caparrini, M., Floury, N., 2014. Airborne GNSS-R soil moisture and above ground biomass observations. IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. 7 (5), 1522–1532.

Elshorbagy, A., Parasuraman, K., 2008. On the relevance of using artificial neural networks for estimating soil moisture content. J. Hydrol. 362, 1–18.

Entekhabi, D., et al., 2014. SMAP Handbook–Soil Moisture Active Passive: Mapping Soil Moisture and Freeze/Thaw from Space.

Entekhabi, D., Reichle, R., Koster, R.D., Crow, W.T., 2010. Performance metrics for soil moisture retrievals and application requirements. J. Hydrometeorol. 11 (3), 832–840. https://doi.org/10.1175/2010JHM1223.1.

Eroglu, O., Kurum, M., Boyd, D., Gurbuz, A.C., 2019. High spatio-temporal resolution CYGNSS soil moisture estimates using artificial neural networks. Rem. Sens. 11, 2272. https://doi.org/10.3390/rs11192272.

Guerriero, L., Martin, F., Mollfulleda, A., Paloscia, S., Pierdicca, N., Santi, E., Floury, N., 2020. Ground-based remote sensing of forests exploiting GNSS signals. IEEE Trans. Geosci. Rem. Sens. 58 (10), 6844–6860.

Hodges, E., Chew, C., Small, E., Al-Khaldi, M., Ouellette, J.D., Johnson, J.T., Lei, F., Kurum, M., Gurbuz, A., Senyurek, V., Xu, X., Shah, R., Yueh, S., Hayashi, A., Setti Jr., P.T., Tabibi, S., Santi, E., Pettinato, S., Roberts, T.M., Colwell, I., Lowe, S., Ruf, C.S., Moghaddam, M., 2024. A blended CYGNSS soil moisture product partitioned with ancillary data. United States National Committee of URSI National Radio Science Meeting, USNC-URSI NRSM 2024 - Proceedings 174. https://doi.org/10.23919/USNC-URSINRSM60317.2024.10464722.

Hornik, K., 1989. Multilayer feed forward network are universal approximators. Neural Network. 2 (5), 359–366.

Linden, A., Kinderman, J., 1989. Inversion of multi-layer nets. Proc. Int. Joint Conf. Neural Networks 2, 425–443.

Liu, Q., Zhang, S., Li, W., Nan, Y., Peng, J., Ma, Z., Zhou, X., 2023. Using robust regression to retrieve soil moisture from CyGNSS data. Rem. Sens. 15, 3669. https://doi.org/10.3390/rs15143669.

Marrs, J., Ni-Meister, W., 2019. Machine learning techniques for tree species classification using Co-registered LiDAR and hyperspectral data. Rem. Sens. 11, 819.

Nabi, M.M., Senyurek, V., Gurbuz, A.C., Kurum, M., 2022. Deep learning-based soil moisture retrieval in CONUS using CYGNSS Delay–Doppler maps. In: IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 15, pp. 6867–6881. https://doi.org/10.1109/JSTARS.2022.3196658.

O'Neill, P.E., Chan, S., Njoku, E.G., Jackson, T., Bindlish, R., 2018. "SMAP L3 Radiometer Global Daily 36 Km EASE-Grid Soil Moisture, Version 5," Boulder, CO, USA. NASA National Snow & Ice Data Center Distributed Active Archive Center. https://doi.org/10.5067/ZX7YX2Y2LHEB.

Pal, M., 2005. Random Forest classifier for remote sensing classification. Int. J. Rem. Sens. 26, 217–222.

Panciera, R., Walker, J.P., Jackson, T.J., Gray, D.A., Tanase, M.A., Ryu, D., Monerris, A., Yardley, H., Rüdiger, C., Wu, X., Gao, Y., Hacker, J.M., 2014. The soil moisture active passive experiments (SMAPEx): toward soil moisture retrieval from the SMAP mission. IEEE Trans. Geosci. Rem. Sens. 52 (1).

Pierdicca, N., Comite, D., Camps, A., Carreno-Luengo, H., Cenci, L., Clarizia, M.P., Costantini, F., Dente, L., Guerriero, L., Mollfulleda, A., Paloscia, S., Park, H., Santi, E., Zribi, M., Floury, N., 2022. Potential of spaceborne GNSS reflectometry for soil moisture, biomass and freeze-thaw monitoring: summary of an ESA-funded study. IEEE Geoscience and Remote Sensing Magazine 8–38.

Prechelt, L., 1998. Early stopping-but when?. In: Neural Networks: Tricks of the Trade. Springer, Berlin, Heidelberg, pp. 55–69.

Quinlan, J.R., 1993. Combining instance-based and model-based learning. In: Proceedings of the Tenth International Conference on International Conference on Machine Learning, pp. 236–243. Amherst, MA, USA, 27–29 June 1993; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA.

Ruf, C.S., Atlas, R., Chang, P.S., Clarizia, M.P., Garrison, J.L., Gleason, S., Katzberg, S.J., Jelenak, Z., Johnson, J.T., 2015. New Ocean winds satellite mission to probe hurricanes and tropical convection. Bull. Am. Meteorol. Soc. 97 (3), 385–395.

Santi, E., 2016. Neural Networks applications for the remote sensing of hydrological parameters. Artificial Neural Networks - Models and Applications Book. InTechOpen. ISBN 978-953-51-2705-5.

Santi, E., Clarizia, M.P., Comite, D., Dente, L., Guerriero, L., Pierdicca, N., Floury, N., 2022. Combining cygnss and machine learning for soil moisture and forest biomass retrieval in view of the ESA Scout hydrognss mission. 2022 IEEE International Geoscience and Remote Sensing Symposium. Kuala Lumpur, Malaysia, pp. 7433–7436. https://doi.org/10.1109/IGARSS46834.2022.9884738.

Santi, E., Paloscia, S., Pettinato, S., Fontanelli, G., 2016. Application of artificial neural networks for the soil moisture retrieval from active and passive microwave spaceborne sensors. Int. J. Appl. Earth Obs. Geoinf. 48, 61–73. https://doi.org/10.1016/j.jag.2015.08.002.

Santi, E., Paloscia, S., Pettinato, S., Fontanelli, G., Clarizia, M.P., Comite, D., Dente, L., Guerriero, L., Pierdicca, N., Floury, N., 2020. Remote sensing of forest biomass using GNSS reflectometry. IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. 13, 2351–2368. https://doi.org/10.1109/JSTARS.2020.2982993.

Santi, E., Clarizia, M.P., Comite, D., Dente, L., Guerriero, L., Pierdicca, N., 2022. Detecting fire disturbances in forests by using GNSS reflectometry and machine learning: a case study in Angola. Rem. Sens. Environ. 270, 112878. https://doi.org/10.1016/j.rse.2021.112878.

Senyurek, V., Lei, F., Boyd, D., Gurbuz, A.C., Kurum, M., Moorhead, R., 2020. Evaluations of machine learning-based CYGNSS soil moisture estimates against SMAP observations. Rem. Sens. 12, 3503.

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., 2012. The Murrumbidgee soil moisture monitoring network data Set. Water Resour. Res. 48, W07701.

Su, Z., Wen, J., Dente, L., van der Velde, R., Wang, L., Ma, Y., Yang, K., Hu, Z., 2011. The Tibetan plateau observatory of plateau scale soil moisture and soil temperature, Tibet - obs, for quantifying uncertainties in coarse resolution satellite and model products. Hydrol. Earth Syst. Sci. 15 (7), 2303–2316.

Su, Z., de Rosnay, P., Wen, J., Wang, L., Zeng, Y., 2013. Evaluation of ECMWF's soil moisture analyses using observations on the Tibetan Plateau. Geophys. Res. Atmos. 118 (11), 5304–5318.

Tolsdorf, T., Strobel, M., Harms, D., 2021. US Department of Agriculture Soil Climate Analysis Network (SCAN) Site 2026 Data, Walnut Gulch #1, Arizona. USDA Natural Resources Conservation Service.

Unwin, M.J., Jales, P., Tye, J., Gommenginger, C., Foti, G., Rosello, J., 2016. Spaceborne GNSS-reflectometry on TechDemoSat-1: early mission operations and exploitation. IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. 9 (10), 4525–4539. https://doi.org/10.1109/JSTARS.2016.26038461.

Unwin, M.J., Pierdicca, N., Cardellach, E., Rautiainen, K., Foti, G., Blunt, P., Guerriero, L., Santi, E., Tossaint, M., 2022. An introduction to the HydroGNSS GNSS reflectometry remote sensing mission. IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. 14 (9456091), 6987–6999. https://doi.org/10.1109/JSTARS.2021.3089550.

Yan, Q., Huang, W., Jin, S., Jia, Y., 2020. Pan-tropical soil moisture mapping based on a three-layer model from CYGNSS GNSS-R data. Remote Sens. Environ. 247, 111947.

Yu, Y., Li, M., Fu, Y., 2018. Forest type identification by random forest classification combined with SPOT and multitemporal SAR data. J. For. Res. 29, 1407–1414.

Zavorotny, V., Gleason, S., Cardellach, E., Camps, A., 2014. Tutorial on remote sensing using GNSS bistatic radar of opportunity. Geosci. Remote Sens. Mag. 2, 8–45. https://doi.org/10.1109/MGRS.2014.2374220.

Zribi, M., Guyon, D., Motte, E., Dayau, S., Wigneron, J.P., Baghdadi, N., Pierdicca, N., 2019. Performance of GNSS-R GLORI data for biomass estimation over the Landes Forest. Int. J. Appl. Earth Obs. Geoinformation 74, 150–158.

Zribi, M., Huc, M., Pellarin, T., Baghdadi, N., Pierdicca, N., 2020. Soil moisture retrieval using gnss-R data. 2020 Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS), pp. 172–175. https://doi.org/10.1109/M2GARSS47143.2020.9105320. Tunis, Tunisia.

IX. Sitography

ESA CCI: https://www.esa-landcover-cci.org/?q=node/164.
GTOPO 30 DEM https://www.usgs.gov/media/files/gtopo30-readme.