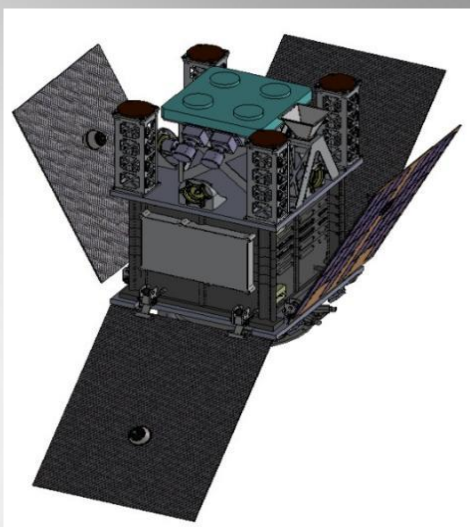
Ref. CRAS-ATBD/6.2025
Issue: 2.0
Rev.: 9.0
Date: 10/06/2025
ESA contract number: 4000135686

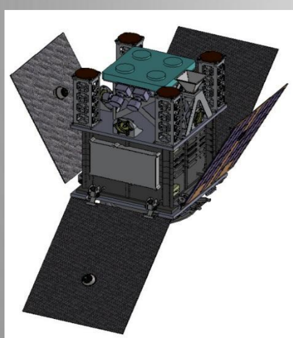
HydroGNSS

Level 2 Soil Moisture Operational Processor v. 3.4 (L2OP-SSM) ATBD

N. Pierdicca, H. Izadgoshasb, F. Cordari

Sapienza University of Rome. Centro Ricerche Aerospaziali (CRAS)





Export Control Document Status

UK Export Control	**This document and any hardware referred to within has been UK Export Control Rated as:**	
Hardware UK Export Rating Number:	N/A Not Controlled	**Date of rating:** 07/07/2025
Information UK Export Rating Number:	N/A Not Controlled	
I confirm I am an approved SSTL Export Rater and this rating is correct as of the date above:		Martin Unwin
US Export Control	No US Content	
ITAR EAR	**This document contains information controlled by the U.S. government and authorized for export only to the country of ultimate destination for use by the ultimate consignee or end-user(s) herein identified. It may not be resold, transferred, or otherwise disposed of, to any other country or to any person other than the authorized ultimate consignee or end-user(s), either in its original form or after being incorporated into other items, without first obtaining approval from the U.S. government or as otherwise authorized by U.S. law and regulations.**	
ITAR	**The information contained herein is controlled under US ITAR and is authorised for export under the provision of TAA/DSP: N/A Not ITAR Controlled**	
EAR	**The information contained herein is controlled under US EAR and has been classified as ECCN: N/A Not EAR Controlled**	

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

DOCUMENT REVISION STATUS

Last Edited Date	Revision / Release Number	Status	Edited By	Pages / Paragraphs Affected	Change Ref
December 7th 2021	1		N. Pierdicca and Sapienza team		
May 20th 2022	2		N. Pierdicca and Sapienza team	Section 3.3 plus minor changes	
October 20th, 2022			N. Pierdicca and Sapienza team	• Reorganization of the document to clarify the different steps for algorithm developments • Description of Level 2G trackwise-gridded ouput added (compared to L2 and L3) • Specification of quality flags added • Clear delineation of Placeholder, Alpha, Beta processors added • Verification and Validation needs added • Auxiliary data consistency checked • L2G product format description • Added details on actual progress in processor development • Explained differences between E2ES and future operational processor in PDGS related to availability of auxiliary data	
January 20th, 2023			N. Pierdicca and Sapienza team	• Update of section 2.1 to better explain roadmap • Updated Fig. 2.1 • Minor changes in 3.6.2 to make the roadmap more clear • Update of section on actual implementation of alfa release with a new AFS parameter model and details on stratification	
October 15th, 2023			Sapienza team	• Definition of beta processor based on ANN • Definition of processing of dynamic auxiliary data (SMAP and MODIS) • Update of Cal/Val section • Removal of the product format section (it is in the User Guide)	
January 2nd, 2024	2.5		Sapienza team	• Correction of typo • Inclusion of an output data section • Change of document referencing • Change of equation numbering	

March 7th, 2024			Sapienza team	Changes to implement actions following the ESA RID • RID-MPC-TRR-6: The eq. 3.12, 3.13, 3.14, 3.5 have been corrected. • RID MPC-TRR-7: In 3.7.2 it is specified the alfa release is assuming a linear relation between R_{dB} and m_v • RID-MPC-TRR-8: specified in 4.1.1 that the alfa release is compliant with the minimum requirements, but it does not achieve the goal • RID-MPC-TRR-9: better explained in 3.1 the coexistence of semiempirical and ANN algo in the PDGS till the Commissioning. • RID-MPC-TRR-10: section 3.7.2 revised to explain how AFS and σ^0 are retrieved. • RID-MPC-TRR-11: the role of RHCP is discussed in section 3.7.3.1 with reference to the simulated experiments. • MPC-TRR-13a and MPC-TRR-12: in section in section 3.1 clarified that both algorithms are in the beta release, they can be selected by a switch, and only the best one will be used during operations.	
November 17th 2024	2.7		Sapienza team	• Updated to match the processor version and according to responses to ESA RID's (actual filtering, flagging, list of input auxiliary data in Table 1, a new reference) • Added Table 3.3 with ANN inputs	
January 25th, 2025	2.8		Sapienza team	• Updated section 3.8 about output quality flags • Specified the signals that are exploited in section 3.7.3	
June 10th 2025	2.9		Sapienza team	• General review to make the doc up-to-date (e.g., refreshing verb tense) • Removed sections referring to functionalities strictly related to E2ES (sections 3.7.1.1, 3.7.2.2, 3.7.3.1) • Updated Applicable Document list	

1 INTRODUCTION

1.1 SCOPE

This document contains the Algorithms Theoretical Baseline Document of the placeholder, alfa and beta releases of the Level 2 processor aimed at generating the HydroGNSS Surface Soil Moisture product as an element of the mission PDGS. The software has an operational mode that allows it to be integrated within the HydroGNSS End to End Simulator (H-E2ES) released at the CDR, which was exploited to assess the system performances before the launch.

The placeholder L2PP was based on a simple purely empirical multivariate regression approach whose main purpose was testing all the E2E software interfaces, inputs and outputs.

The Level 2 alfa Operational Processors (L2OP) implements a physically based retrieval algorithm based on the set of observables delivered by HydroGNSS, as well as static and dynamic auxiliary data.

The Level 2 beta release exploits an advanced Artificial Neural Network (ANN) data driven approach exploiting also HydroGNSS and a more extensive set of static and dynamic auxiliary data.

During the algorithm development activities, before the launch, the physically based and ANN algorithms will be compared in terms of performances, i.e., both thematic accuracy of soil moisture estimates and computational time before making the final decision about the approach to be exploited during operations.

1.2 APPLICABLE DOCUMENTS

Applicable Documents are identified by AD-n, where “n” indicates the actual document number, from the following list:

AD#	Title	Doc #	Revision	Date
1	HydroGNSS PDGS Algorithm ICD	00637207	17	03/03/2025
2	Scout-2 HydroGNSS SRD Science Team Annex June 2021	00635466	3	01/10/2021
3	HydroGNSS L2 Placeholder Processor ATBD0	CRAS-ATBD/1.2021	1	05/12/2021
4	Scout-2 HydroGNSS - Mission Requirements Document	00632438	1.5	12/8/2021
5	HydroGNSS Soil Moisture Level 2 Beta Processor (SM-L2OP) V3.3 User Guide	CRAS-L2OP/3.2025	3.3	25/01/2025
6	Level 2 Soil Moisture Operational Processor v. 3.3 (L2OP-SSM) ATBD	CRAS-ATBD/1.2025	2.8	25/01/2025
7	HydroGNSS E2ES Algorithm Theoretical Baseline Document	CRAS-E2E/2.2024	2.1	20/06/2024

1.3 ACRONYMS AND ABBREVIATIONS

The following abbreviations are used within this document:

AFS	Attenuated Forward Scattering
AGB	Forest Aboveground Biomass

ANN	Artificial Neural Networks
ANN	Artificial Neural Network
ATBD	Algorithm Theoretical Baseline Document
CCI	ESA's Climate Change Initiative
CGR	Completion Gate Review
DCM	Document Completion (internal) Milestone
DDM	Delay Doppler Map
DEM	Digital Elevation Model
DGR	Delivery Gate Review
DTM	Digital Terrain Model
E2ES	End To End Simulator
EASE	Equal-Area Scalable Earth
EIRP	Effective Isotropic Radiated Power
GNSS	Global Navigation Satellite System
GPS	Global Positioning Satellites
GTOPO 30	USGS 30 ARC-second Global Elevation Data
H-E2ES	HydroGNSS End To End Simulator
H-SAVERS	HydroGNSS SAVERS
IFAC	Institute of Applied Physics
L1	Level 1
L2	Level 2
L2G	Level 2 gridded
L2OP	Level 2 Operational Processor
L2OP-SSM	Level 2 Operational Processor for Soil Moisture
L2PP	Level 2 Placeholder Processor
L3	Level 3
LE	Leading Edge
LHCP	Left Hand Circular Polarization
MMSE	Minimum Mean Square Error
MODIS	Moderate Resolution Imaging Spectroradiometer
NBRCS	Normalized Bistatic Radar Cross Section
NDVI	Normalized Differential Vegetation Index
NDWI	Normalized Differential Water Index
NN	Neural Network
PAM	Performance Assessment Module
PDGS	Payload Data Ground Segment
PR	Polarimetric Ratio
RHCP	Right Hand Circular Polarization
RMS	Root Mean Square
RMSE	Root Mean Square Error
SAVERS	Soil And VEgetation Reflection Simulator
SCM	Software Completion (internal) Milestone
SGR	Status Gate Review
SM	Soil Moisture
SMAP	Soil Moisture Active Passive

SMOS	Soil Moisture Ocean Salinity
SP	Specular Point
SRTM	Shuttle Radar Topography Mission
SSM	Surface Soil Moisture
SSTL	Surrey Satellite Technology Limited
TBC	To be considered
TBD	To be defined
TBV	To be verified
TE	Trailing Edge
ubRMSE	Unbiased RMSE
VOD	Vegetation Optical Depth

2 BACKGROUND AND SCIENTIFIC JUSTIFICATION

2.1 OVERVIEW OF THE SOFTWARE ENVIRONMENT

The Level 2 (L2) processor is a component of the H-E2ES, but it is also a component of the PDGS during the operational phase of the mission. Figure 2-1 shows the overall structure of the H-E2ES and the collocation of the L2 processors in such architecture. The main interface among different modules coincides with the products itself delivered by the mission at different levels. Besides this, the L1b and L2 modules interfaces with the auxiliary data, as they are stored in the ground segment. CRAS, Sapienza University of Rome, is responsible for the H-SAVERS and the L2OP-SSM module producing the soil moisture product (one of the green boxes in the figure).

The L2 processors were released in three different steps. The first release ("placeholder") was delivered in time to be integrated within the H-E2ES at the PDR. At PDR, H-E2ES simulated only 1st order types of observations from HydroGNSS, as implemented in previous missions like TDS-1 and CyGNSS (i.e., it was not foreseen to simulate new observables such as the complex channel and dual polarization). Then the "placeholder" L2 processor (L2PP) only considered those observables in input. After the PDR, the 2nd order types of observables were simulated by the H-E2ES, so that the L2 operation processors were updated at CDR to accommodate those observables when possible (L2OP alfa release). Finally, the beta algorithms were developed at TRR, and the L2OP beta release processors are placed side by side within the PDGS, as well as integrated in the H-E2ES.

AS far as the HydroGNSS E2ES (H-E2ES) is concerned, it integrates the following elements (see Figure 2-1):

- SAVERS, updated (H-SAVERS) to account for other targets and instrument effects (contributions from IEEC Wavpy were provided in the form of ATBD and/or Matlab© scripts, including the simulation of the complex channel and of correlated speckle noise in the DDM)
- L1A to L1B processor L1PP (3 releases: placeholder at PDR, α- at CDR, β- at TRR)
- L1B to L2 processors L2PP's (3 releases: placeholder at PDR, α- at CDR, β- at TRR)

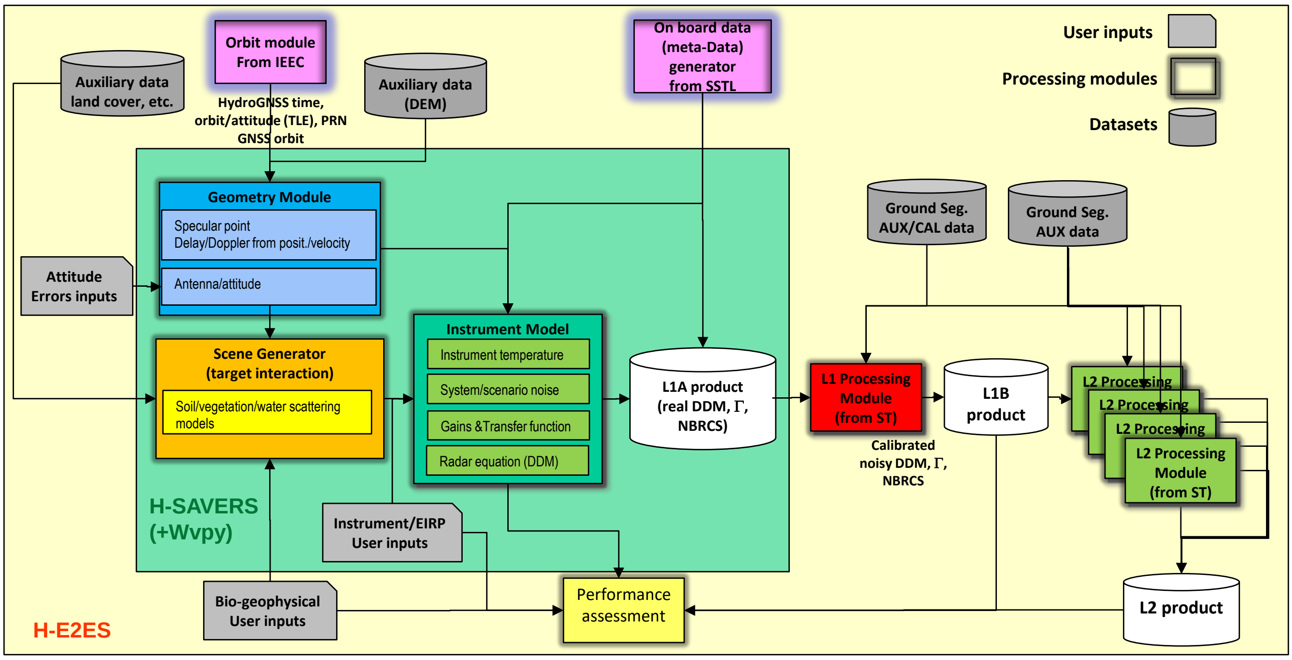


Figure 2-1. Top Level Functional Diagram Structure of the HydroGNSS E2ES and its interfaces.

The placeholder L2PP is based on a simple purely empirical multivariate regression approach, whose main purpose was testing all the E2E software interfaces, inputs, and outputs. The L2OP alfa release is a model-based algorithm; it exploits HydroGNSS data inverting a semi-empirical forward model and using auxiliary data, both static and dynamic, to account for confounding parameters affecting the observations (e.g., the topography or the changing vegetation conditions). The beta release adds an advanced Artificial Neural Network (ANN) data driven approach taking as inputs HydroGNSS and a larger set of auxiliary data.

The software development was incremental, so the beta release implements both types of algorithms that can be selected through the configuration file. During the algorithm development activities and later during the Commissioning Phase the model-based and ANN algorithm performances can be compared in terms of both thematic accuracy and computational time before taking the final decision about the approach to be used operationally in the PDGS. As for the coefficients of the algorithms, they are stored in an algorithm configuration file, so that they can be easily updated considering new training datasets. The training sets were generated by the H-E2S itself simulating the signal over a number of sites. To extend the algorithm capability on a global scale, CyGNSS data were the main source of training set, at least for the conventional type of measurements (i.e., DDM's in power units at LHRP).

It must be pointed out that a difference exists between simulated and real GNSS-R data for what concerns the dynamic auxiliary information. In case of real data, such as CyGNSS and HydroGNSS, the dynamic auxiliary information about the targeted area can be provided by other satellite missions observing the same site. When simulated data are used, the random realizations of the soil and vegetation confounding parameters assigned by the simulator cannot be provided by any real observation; they must be passed by the H-SAVERS L1a simulator to the L2OP processor. A switch in the configuration file is present to run the processor within the H-E2ES with simulated data or within the PDGS with real data.

2.2 REVIEW OF SURFACE SOIL MOISTURE FROM GNSS-R

The peak power of a GNSS-R DDM has been shown to be sensitive to the dielectric constant (a proxy for soil moisture over bare soil) and surface roughness. The signal is also affected by the amount of vegetation and its water content and salinity (Ulaby and El-rayes 2007). Such results have been demonstrated by ground based and airborne experiments, but also by satellite data using both TDS-1 and CYGNSS DDMs e.g. (Chew, Shah, et al. 2016). Notably, forward near specular reflection is directly related to soil permittivity (as for a monostatic radar) and inversely related to roughness and vegetation optical depth (contrary to the monostatic case). Then, besides the inherent sensitivity to soil moisture, this highlights also a role of bistatic GNSS-R measurements to complement monostatic radar and provide synergic information.

Early works demonstrated the sensitivity of the GNSS reflection to soil moisture (Masters 2004). Ground based and airborne data were collected in two ESA experiments investigating agricultural and forestry applications (Egido, Caparrini, et al. 2012), (Egido, Paloscia, et al. 2014) using a receiver with dual polarisation.

From TDS-1 satellite data, comparisons between GNSS-R SNR and SSM data have shown a large sensitivity (amount of dB's for a change of soil moisture) of about 38 dB/m^3/m^3 for bare soils and a high Pearson correlation coefficient R =0.63 (Camps, Park, et al. Oct. 2016). However, this sensitivity rapidly decreases as the vegetation grows (NDVI was used as a global vegetation proxy), the impact of vegetation being mainly an attenuation of the transmitted signals related to the vegetation optical depth (VOD). Using airborne GNSS-R data, in (Sánchez, et al. 2015) it was found that, after correcting for variations of the local incidence angle (i.e. topography effects), the best results were obtained combining the Green-Red Vegetation Index from Landsat 8 and the GNSS-R reflectivity in the 40°-60° incidence angle range. This demonstrates the importance of ancillary data in the retrieval. The correlation between TDS-1 reflectivity and soil moisture delivered by SMOS was investigated in (Camps, Vall-Llossera, et al. 2018), where the sensitivity (dBs for a change of 10% in soil moisture, which corresponds to 0.1m^3/m^3) was assessed in the order of 0.7 dB/10% after correcting the vegetation effect using VOD data. It was observed that the sensitivity decreases for larger incidence angles (smaller elevation of the transmitter) and the effect of the vegetation cover is difficult to compensate. It is interesting that in the same project the sensitivity evaluated using CYGNSS data compared to SMAP SM, after a careful filtering of good quality data, was 1.1 dB/10%, fairly similar, and even better, to the previous case.

In (Chew, Shah, et al. 2016) the authors demonstrate the sensitivity of GNSS-R measurements to soil moisture using TDS-1 L1 DDMs. Their method uses the effective peak power of a DDM ($P_{r,eff}$) calculated from the highest reflected power of each DDM ($P_{r\ dB}$) and various noise and geometry terms, described by the following equation:

$$P_{r,eff} \propto P_{r\ dB} - N_{dB} + (R_{sr} + R_{ts})^2_{dB} - G_{r,dB} + \cos^2\theta$$

N corresponds to the receiver's noise, calculated using the noise box method. R_{sr} is the specular reflection point to receiver range, R_{ts} is the transmitter to specular reflection point range, θ is the incidence angle, and $G_{r,dB}$ is the antenna gain toward the specular reflection point in dB.

On August 2015, the radar receiver aboard the SMAP satellite was tuned to detect GNSS reflections in the GPS L2C signal band, at horizontal and vertical polarisations. An investigation of those data and the potential of a dual polarisation system (linear polarisations in this case) was carried out in (Carreno-Luengo H., 2017). In this work, the Polarimetric Ratio (PR) showed a good sensitivity to soil moisture content over different land cover types, thus confirming the findings in (Egido, Caparrini, et al. 2012). PR showed a correlation coefficient around 0.6 to soil moisture content by comparing it to the SMAP radiometer data.

More recently, the better coverage of the tropical latitude belt of the CYGNSS constellation strongly stimulated the development of soil moisture products, enabling the comparison of spatial patterns of reflectivity aggregated according to different grid scale and soil moisture maps from SMOS and SMAP.

Note that UCAR/CU CYGNSS presently delivers an experimental Soil Moisture Product at a global scale, combining CYGNSS and SMAP data, which has been demonstrated to be quite effective in densifying the SMAP SM and accurate when compared to in situ data (Chew and Small 2018), (Chew C. 2019b). This is of course limited to the tropical and equatorial area covered by the CYGNSS constellation, whereas HydroGNSS will cover almost any latitude.

Further algorithms using CYGNSS data have been developed recently, showing very encouraging results. A change detection algorithm exploiting time-series of CYGNSS normalized radar cross section, assuming purely incoherent signal, was developed in (Al-Khaldi M. 2019). The method requires prior information to solve the inversion problem and for this purpose it exploits maximum and minimum monthly soil moisture from SMAP at each location. Products were generated on a grid of 0.2° X 0.2° and on daily and 3-day bases. The comparison with SMAP products showed an RMSE in the order of 0.040 cm^3/cm^3 and a correlation around 0.82.

(Clarizia, et al. 2019) introduced a trilinear regression exploiting CYGNSS reflectivity only (i.e., without relying on other SM products) and auxiliary data from SMAP (vegetation optical depth and roughness) to derives daily SM estimates at a 36 km X 36 km resolution. The algorithm was developed considering the dominance of the coherent reflections over land. The performance was compared globally against an independent set of SMAP SM products and reported to have a RMSE of 0.07 cm^3/cm^3, which reduces to around 0.06 cm^3/cm^3 if the regression is trained at regional scale.

In (Eroglu 2019) an artificial neural network method was trained and tested using in situ data collected in 18 International Soil Moisture Network (ISMN) sites, considering a box of 9km around each in situ sample. The method combines observables from CYGNSS (e.g., reflectivity, trailing edge slope, incidence angle) and ancillary data (e.g., NDVI, surface height and slope), but it does not relay on other SM products. It aims at fully exploiting the spatial resolution of CYGNSS and provides sub-daily products. The product validation resulted in an unbiased root mean square error (ubRMSE) of 0.05 cm^3/cm^3 and a correlation coefficient of R=0.9.

During the HydroGNSS Consolidation Study an algorithm for SM retrieval from CYGNSS reflectivity based on an ANN approach and integrating auxiliary data has been trained and tested using SMAP products (only 1% of SMAP data used for training and the rest used for testing). The test resulted in a RMSE of 0.06 m^3/m^3. The beta processor described in the next sections is rooted on that work.

Neither the TDS-1 nor CYGNSS missions were conceived to sense land, as the instruments were primarily designed for marine applications. Key to the current soil moisture retrievals is the calibration of the L1 DDMs accounting for receiver antenna gain, bistatic geometry and transmitter EIRP and considering the complex nature of the reflection from land, combining coherent and incoherent contributions (Chew, Small and Read 2019).

2.3 COHERENT AND INCOHERENT SIGNAL

An important concept to point out here is the nature of the scattered signal collected around the specular direction that can be coherent in case of a perfectly flat surface or purely incoherent in case of a very rough surface (e.g., the ocean surface with big waves). In a real situation the signal is neither fully coherent nor fully incoherent, but rather partially coherent. This is a topic still under discussion in the scientific community, and many researchers developed works to quantify the degree of coherence from the data itself (raw IF data or detected DDM's). In the two extreme cases, the quantity characterizing the target are the reflectivity Γ or the Normalized Bistatic Radar Cross Section (NBRCS).

The reflectivity of a flat surface is defined by equation (2.1) assuming the radar equation for a coherent signal, so that the received power P_r^S (i.e., the one at the receiver input) coherently reflected by the surface coincides with only its coherent component P_c^S and is given by:

$$P_r^S = P_c^S = \Gamma(\theta)\frac{\lambda^2 G^t\, G^r\, P_t}{(4\pi)^2(R_r+R_t)^2} \tag{2.1}$$

Where P_t is the transmitted power, R_r is the distance from receiver to the specular point (SP), R_t that from transmitter to SP, λ is the wavelength, G^t and G^r are the gain of transmitting and receiving (i.e., down-looking) antennas. Γ is the surface reflectivity, which is a function of the incidence angle θ at the SP. In case of a flat surface $\Gamma(\theta)$ is equal to the square module of the Fresnel reflection coefficient $|R|^2$. The signal power P_r^S is assumed being derived from the received power subtracted by an estimate of the noise additive power. The signal received by the down-looking antenna is partially generated by incoherent scattering within the antenna footprint, with the relative contribution from coherent and incoherent depending on the receiver height, roughness and eventually volume attenuation and scattering. Therefore, in a real case we can still refer to this definition, but we will speak about an "equivalent" reflectivity which account for both contributions.

The Normalised Bistatic Radar Cross Section (NBRCS) or Bistatic Scattering Coefficient (σ°) is introduced in case of incoherent scattering, when the received power equals its incoherent component P_i^S, and is defined by:

$$P_r^S = P_i^S = \sigma^o(\boldsymbol{i},\boldsymbol{s})\frac{\lambda^2 G^t\, G^r\, P_t A_S}{(4\pi)^3 R_r^2 R_t^2} \tag{2.2}$$

Where A_s is the illuminated area of the resolution cell determined by the range and Doppler discrimination of the system and σ° is a function of the incidence direction **i** and observation directions **s** of the resolution cell. The distances are those from transmitter and receiver to the surface element A_s. The meaning of the other quantities is the same of previous equation. Also in this case, if we have some amount of coherent reflection, we can stick on this definition, but we can refer to an "equivalent" NBRCS.

Considering the discrimination mechanism of a GNSS-R system, the previous formula can be generalized as the scattered signal comes from the entire footprint of the antennas (intersections of receiving and transmitting antenna) and the contributions from the different cells are modulated by the Woodward ambiguity function χ^2 depending on the delay time τ and Doppler frequency f_D. Thus, the following bistatic radar equation applies:

$$P_r^S = \frac{\lambda^2 P_t}{(4\pi)^3}\int \frac{G^t\, G^r}{R_r^2 R_t^2}\sigma^o(\mathbf{i},\mathbf{s})\chi^2(\tau, f_D)dA \tag{2.3}$$

Where distances go from transmitter and receiver to the generic point on the integration surface.

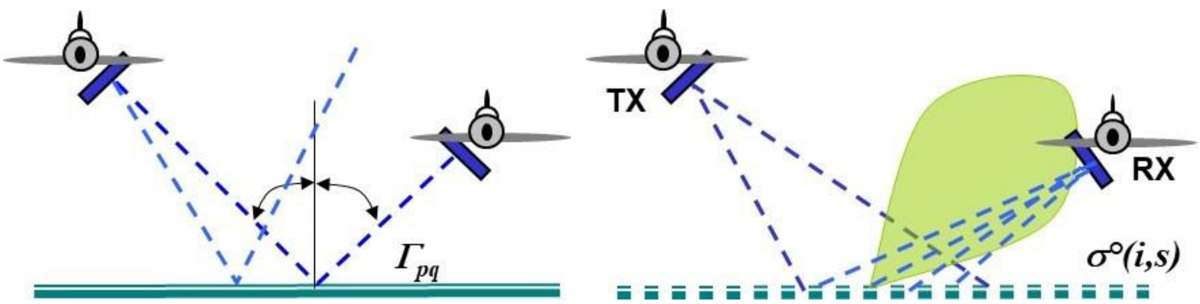


Figure 2-2. Graphic representation of the coherent reflection from a flat surface (left) and diffuse incoherent scattering from a random surface

The adjective "equivalent" considers that for each case there is also a certain amount of the other scattering mechanisms contributing to the received power. In Figure 2-2 a sketch explaining the two different mechanisms is presented. Note that the signal received power, the effective reflectivity and NBRCS carry equivalent information and can be converted into the others by formulas involving only geometric quantities. Note that over a land surface with rough topography the computation of the area of the surface effectively illuminated within a certain doppler and range resolution cell can be difficult.

2.4 OVERVIEW OF SSM ALGORITHMS FOR GNSS-R

The mostly used GNSS-R observable over land is the "equivalent" reflectivity Γ, which is defined considering the signal is dominated by the near specular reflection, as done for instance in (Clarizia, et al. 2019). Alternatively, one can consider the Normalized Radar Bistatic Radar Cross Section (NBRCS), which models the target in case of a completely incoherent signal. This is typically used over ocean.

Regarding the Level 2 processor for SSM (L2OP-SSM) the retrieval baseline algorithms to be implemented have been reviewed during the Consolidation Study and advantages and drawbacks have been indicated. At present, four approaches can be foreseen for land application retrievals, based on the experience gained from the literature and from two ESA studies: "GNSS over land" (ESA Contract 4000120299/17/NL/AF/hh) and "Exploiting CYGNSS GNSS-R over Land and Ocean for GeophYsics (ECOLOGY) (ESA Contract 4000130560/20/NL/FF/gp/).

- SM model-based inversion. In this case, the reflectivity of the canopy Γ (soil and vegetation) is expressed as a function of the reflectivity of the flat soil Γ^{flat}, diminished by the small-scale roughness factor and by the vegetation attenuation as discussed e.g. in (Camps, Cenci, et al. 2019) (Pierdicca et al., 2022) (Camps, Vall-Llossera, et al. 2018). This approach assumes the observed signal is purely coherent so that the following formula can be written:

$$\Gamma = \Gamma^{flat} \cdot e^{\left(-4k^2\sigma^2 cos^2(\vartheta)\right)} \cdot e^{\left(\frac{-2\tau}{\cos(\vartheta)}\right)} \tag{2.4}$$

  Where k is the electromagnetic wavenumber, σ is the soil roughness standard deviation, τ is the vegetation optical depth (VOD) and θ is the incidence angle (with respect to the vertical). By inverting the model, the reflectivity of the flat soil can be related to the permittivity through the Fresnel reflection coefficient formula and then to the soil moisture. The main advantage of the model belongs to its theoretical formulation that does not need empirical coefficients and calibrations. Conversely, the main drawbacks of the approach are the need to collect accurate ancillary data, and its simplified formulation that disregards the incoherent contribution and volume scattering of vegetation. The use of proxies for ancillary data, for instance the use of the Normalized Differential Vegetation Index (NDVI) instead of VOD, and the issues related to the different spatial resolution of GNSS-R and auxiliary data can heavily affect the model predictions and the inversion can lead to large errors in estimating SM.

- SM semi empirical regression. This approach has been proposed by (Clarizia, et al. 2019) and it can be regarded as a simplified data-driven implementation of the model-based inversion. When considering the reflectivity in dB, its relationship with roughness and opacity becomes linear and it can be expressed as:

$$\Gamma^{flat}{}_{dB} = a \cdot \Gamma_{dB} + b \cdot \tau + c \cdot \sigma + d \tag{2.5}$$

  where the model coefficients are derived through a multivariate linear regression. The Fresnel reflection coefficient is then assumed linearly related to soil moisture. The main advantage is the simplicity of the approach that is easy to implement. However, it needs auxiliary information about roughness and vegetation conditions, as for the model approach, and it is based on a training phase to define the empirical coefficients. In (Clarizia M.P. 2019) the model coefficients were derived by a training set through a tri-variate linear regression; SMAP vegetation opacity and roughness coefficients were considered as ancillary information, all of them posted on the same space and time grid. In the same category one can consider the algorithm developed by UCAR/CU that delivers a quasi-operation SM product from CyGNSS (Chew C. 2019b) (Clara Chew 2020). Opacity and roughness effects are not considered as the regression coefficients are updated continuously using the product delivered by SMAP. The empirical regression approaches can be useful to adjust the forward model parameters, to accommodate ancillary data that may be not accurate enough or not providing exactly the roughness or VOD quantities. However, both spatial and temporal resolution and the quality of the ancillary data remain critical aspects for ensuring successful retrievals.

- SM Change detection. An algorithm that falls into this category was proposed for processing GNSS-R data in (Camps, Cenci, et al. 2019). For each point (or cell of a grid) a different equation is set from historical time-series of data, by identifying the minimum and maximum values of Γ of the given pixel that are then associated to the drier and wetter soil conditions. SM, or at least a soil saturation degree, is linearly related to Γ within this interval. This is the approach that has been already adopted for generating the ASCAT SM products and later SM product from Sensintel-1 (Wagner et al., 1999; Zribì et al., 2021). It is effective and it does not need ancillary roughness and vegetation information since the effect of both is implicitly accounted for by deriving site dependent relationships. However, it needs large datasets for calibrating the algorithm and setting up the Γ minimum and maximum values. Temporal changes of roughness and vegetation conditions may also have a detrimental effect. The exploitation of a time series was also proposed by (Al-Khaldi M. 2019). The method leverages on the slowly changing nature of vegetation and surface roughness compared to soil moisture and derives the change in SM by computing the ratio of normalized bistatic cross section of the incoherent component of the signal between successive acquisitions. It needs some prior information about the statistics of the SM in the area (e.g., minimum and maximum values) that in (Al-Khaldi M. 2019) were derived from SMAP ancillary soil moisture information, a limitation that must be overcome. All these change detection approaches cannot account for any abrupt roughness or vegetation variation within the given cell.

- SM and biomass Artificial Neural Networks. ANN enables data driven (training with experimental data), model driven (training with model simulations) or mixed (simulated + experimental) trainings, thus combining the advantages of experimental-based and of model-based retrievals. The ANN training can have a significant computational cost; however, training must be done only once, and the trained ANN can be saved and applied to a new data dataset in near real time. The main disadvantage of this technique is that the ANN are prone to the so-called outliers: the retrieval can be unsuccessful if the ANN is applied to data falling outside the range of values included in the training set. An example of ANN application for mapping AGB and tree height at global scale can be found in (Santi, et al. 2020), whereas an ANN to retrieve soil moisture has been proposed in (Eroglu 2019).

All the considered algorithms rely on auxiliary data to account for factors affecting the signal to be distinguished from the target geophysical quantity. A prior knowledge of the terrain elevation, small scale roughness and vegetation has to be provided as a minimum requirement. Other information about land use land cover and soil texture and porosity may also be useful to tune the algorithms over different surfaces.

HydroGNSS has additional observables that are not yet available from satellite with respect to the standard LHCP DDM's collected from GPS transmitters. We refer in particular to the dual polarisation capability (detection of circular left and circular right reflected signal) and the complex channel. They can enable to disentangle the effect of concurrent parameters on the signal and in particular the combined effect of roughness and moisture, or moisture and vegetation, to improve retrievals.

3 DESCRIPTION OF THE L2OP-SSM ALGORITHM (ATBD)

3.1 Algorithm development strategy

The algorithm development has progressed through three main phases. Three versions of the processor have been implemented: placeholder, alfa and beta releases. Note that the processors embedded in the E2ES and the one in the PDGS working on real HydroGNSS data have necessarily some differences for what concerns the auxiliary data to be exploited. In fact, the PDGS processor during operation will be able to exploit dynamic auxiliary data (e.g., MODIS, see section 3.3), which cannot be collected within the E2ES since the bio-geophysical parameters of the scenarios are set randomly. For this reason, the auxiliary data within the E2ES can be gathered from the "Bio-geophysical User Input" input shown in Figure 2-1. The actual auxiliary parameters used in the simulation are stored in a Reference File, which is also used by the Performance Assessment Module (PAM).

- The placeholder L2PP is based on a simple purely empirical multivariate regression approach whose main purpose is testing all the E2E software interfaces, inputs and outputs.
- The Level 2 alfa Operational Processors (L2OP) implements a physically based retrieval algorithm based on the observables delivered by HydroGNSS. It was calibrated at the time of the CDR using simulated products from the H-E2ES. In parallel, a preliminary calibration of the algorithm coefficients and some auxiliary data needed (see section 3.7.2) was carried out using CyGNSS and SMAP data to prepare the global scale processing of HydroGNSS data in the PDGS. This calibration is being prosecuted as part of the pre-launch and post-launch cal/val activity.
- The beta release further adds an advanced Artificial Neural Network (ANN) data driven approach. The ANN training within the H-E2ES was done using simulated data, but this led to some questionable results in terms of performances when the size of the simulated dataset was limited (e.g., in case of Galileo signal). The ANN training was also carried out using CyGNSS, SMAP and MODIS data to prepare the global scale processing of HydroGNSS data in the PDGS. It is being continued during pre- and post-launch cal/val activities.

The algorithm development activity is a process that continues almost for the entire lifetime of a mission. The beta release (thus the processor in the PDGS) can run either the semi-empirical or the ANN by just setting an input parameter in the algorithm configuration file. The coefficients of the two algorithms can also be changed by changing the algorithm configuration file. In this way it is possible to compare the two approaches in terms of performances (both thematic accuracy and computational time), provided reliable validation data will be made available. If one of them will not be considered useful during the prosecution of the cal/val activity before launch, only one can be activated. At this stage, as it will be reported in the sequel, the ANN seems to provide better results. It is not foreseen that during the operational phase (after commissioning) the L2 product will be produced twice, as only the 'winner' algorithm will be used at the end.

As for the algorithm coefficients (e.g., weights of each ANN connection), they were initially tuned using the H-SAVERS outputs for assessing the performances using the PAM. In parallel, and in view of implementing the PDGS and generating product at global scale, the training was carried out using as main source of training data the CyGNSS L1B products combined with other auxiliary data (e.g., SMAP). This will prosecute as part of the general cal/val activity. Table 3-1 summarizes the main differences among the processors and the auxiliary data used in inputs. The table includes also suggestions for additional auxiliary data to be considered in the future for improving the performances but not implemented at this stage and before Commissioning; they are denoted as To Be Considered (TBC) in the table.

Table 3-1. Overview of difference across the SSM L2 processors in the frame of E2ES and operational version (PDGS) working at global scale

	L2PP placeholder	L2OP alfa		L2OP beta	
		In E2ES	In PDGS	In E2ES	In PDGS
observable	Reflectivity LR	Reflectivity LR+RR (depends on SNR)	Reflectivity LR+RR (depends on SNR)	Reflectivity LR+RR (depends on SNR)	Reflectivity LR+RR (depends on SNR) Full DDM (TBC)
algorithm	Multivariate regression	Semi-empirical physically based algorithm		Artificial Neural Network	
Small scale roughness	none	From Reference File	AFS map from CyGNSS+SMAP (see 3.7.2)	From Reference File	None AFS if extended to all latitudes from HyDROGNSS +SMAP (TBC)
Topography	From static DEM	From static DEM	From static DEM	From static DEM	From static DEM
Vegetation	none	From Reference File	From dynamic aux (SMAP, MODIS)	From reference file	From dynamic aux (SMAP, MODIS)
Texture	none	none	From static maps (TBC)	none	From static maps (TBC)

Further dynamic auxiliary (to filter out data)	none	none	none	none	Snow data, freeze/thaw state (the latter TBC)
stratification	none	Land cover	Land cover + geolocation	none (land cover is an input)	none (land cover is an input)
Spatial resolution	25 km	25 km	25 km	25 km or less (TBC)	25 km or less (TBC)
Temporal resolution	Track-wise (L2G)	Track-wise (L2G)	Track-wise (L2G) + daily	Track-wise (L2G)	Track-wise (L2G) + daily (L3)

3.2 Track-wise gridded (L2G) and accumulated gridded (L3) products

If the retrieval algorithm is applied to every sample of the L1B product, thus along an individual specular point track sampled every second (this is the sampling rate foreseen for L1B data) we denote the output as a Level 2 product (i.e., track-wise).

However, in this document we will describe a soil moisture product generated on a gridded base. This is needed to mitigate the fluctuations of the reflections observed in previous studies when the signal is predominantly incoherent and/or even the near-specular reflection fluctuates. Thus, averaging within each grid cell can improve the final radiometric resolution. Moreover, the random location of the specular points, typical of the GNSS-R technique, is partially responsible for that fluctuation since observations can never be ascribed exactly to the same resolution cell over the Earth surface. Averaging reflections always within the same cell partially mitigate this problem, especially if a constellation enables a good coverage of the surface and accumulation of a larger number of reflections within the same resolution cell become possible.

The projection used is the Equal-Area Scalable Earth (EASE) grid version 2 coordinate global reference system (EASE-Grid 2.0). The Equal-Area Scalable Earth (EASE) Grids are intended to be versatile formats for global-scale gridded data, including remotely sensed data. Data can be expressed as a digital array with many possible grid resolutions, which are defined in relation to one of four possible projections: Northern/Southern Hemispheres (Lambert's equal-area, azimuthal), temperate zones (cylindrical, equal-area), or global (cylindrical, equal-area) (see https://nsidc.org/ease/ease-grid-projection-gt). As for the grid spacing (resolution) the baseline value is D= 25 km. Finer resolution can be set by a parameter in the processor configuration file (12.5 km and 9 km) but its use will be considered at a later stage of the project, for instance when a reduced incoherent integration time will be proved effective.

Besides the spatial aggregation of observations within a grid cell, one can eventually consider a temporal aggregation along a prescribed time interval. In case a temporal aggregation is not performed, the product is extracted from an individual specular point track and only aggregated spatially within the prescribed grid cells. The product in this case is referred as Level 2G and a near real-time processing is feasible, i.e., the product can be generated as soon as the single track is acquired.

In case of a temporal aggregation the accumulation will span a period of the order of days and observations along different specular point tracks collected by different platforms of the constellation can be merged (thus with different incidence angles). In this case it is foreseen to release a global product typically every day (although not fully covering the Earth surface with one or two satellites), or more days if users prefer a wider coverage (maximum three days so as to match the revisit time of SMAP and SMOS). This is typically a Level 3 product.

As for timeline, L2G product can be generate by direct processing of track acquisitions, then for each six hours block of L1B products. Instead L3 products will be generated on a Nd days base (the baseline is Nd=1 day and not greater than 3 days).

3.3 INPUT DATA

A full list of the foreseen inputs to the L2OP-SSM processor is reported below. The actual inputs exploited by the different releases (e.g., placeholder, alfa and beta) will be listed in the relevant sections.

3.3.1 *HydroGNSS L1b data*

The following pieces of information should be available as a minimum in the L1B products (as for the actual names of these parameters we refer to AD.1):

- Time tag
- Calibrated DDM's (e.g., signal at the antenna port in power units, as a function of delay and Doppler with noise power removed)
  - L1 LHCP only as for placeholder
  - L1 LHCP and RHCP signal as for alfa release (RHCP can be affected by low SNR)
  - L1 and E1 LHCP and RHCP signal as for beta release (RHCP can be affected by low SNR)
- Estimated noise power (at the antenna port)
- Estimated SNR
- Latitude and longitude of the specular point. It should be computed considering the mean elevation of the Earth surface using a DEM
- Incidence angle at the specular point. It should be computed considering the mean elevation of the Earth surface
- L1B Quality flags (see L1B product description)
- Estimated degree of coherence (from complex signal or DDM)
- Transmitter identification (PRN)
- Frequency band and signal modulation ID
- Peak ("equivalent") reflectivity
- Normalized Bistatic Cross Section (NBRCS)
- Transmitter and receiver position and velocity (in ECEF)
- Processing parameters (integration times, DDM size, etc)
- Land cover / land use label of each observation (if available from L1B processor, otherwise taken for a static auxiliary dataset)

3.3.2 *Auxiliary data*

Auxiliary data to be collocated with HydroGNSS data are first reprojected in the same Equal-Area Scalable Earth (EASE) grid version 2 coordinate global reference system (EASE-Grid 2.0).

We can distinguish between "static" data, i.e., updated on a long temporal scale (say months or years), and "dynamic" time-variant data, i.e., available within a time difference order of few days or weeks from the HydroGNSS overpass. The selection of static or dynamic data depends on the availability and reliability at the time of HydroGNSS operations. A general classification of the needed data with their minimum requirements is done hereafter. The datasets currently used in the processor are also reported.

Static

- Digital Terrain Model (DTM). Digital Elevation Model (DEM) of the surface with respect to the WGS84 ellipsoid at a resolution better than 1 km. From this, other terrain model parameters are derived off-line at coarser resolution, within a resolution cell of 25 km or better, to be used in the retrieval algorithm and/or to generate quality indices:
  - mean elevation,
  - elevation standard deviation,
  - mean slope
  - slope standard deviation

  Currently it is exploited the USGS EROS GMTED2010 250m resolution DEM.
- Land use / land cover map. It shall include at least water bodies (river, lakes), ocean, forest, urban areas, ice. This is a static dataset and can be derived from different sources after reduction of resolution, when needed, by computing the modal value within 25 km or better, and aggregating

land covers that are similar for what concerns GNSS-R response (thus reducing the legend complexity). The thematic map is used for stratifying the observation set and apply the right algorithm to each stratum for the semiempirical algorithm, and as an additional input to the neural network algorithm; it is also used to generate output product quality indices.

Currently it is derived from:

- o CCI land cover map ESACCI-LC-L4-LCCS-Map-300m-P1Y-2015-v2.0.7b.nc, volume is 2.5 GB full resolution (this is the baseline) (*https://www.esa-landcover-cci.org/*)

- Soil texture and soil porosity maps, e.g., Global Land Data Assimilation System (GLDAS) (*https://ldas.gsfc.nasa.gov/gldas*). Soil permittivity and specular reflectivity are weakly related to texture; soil porosity determines the ability of soil to store water. They can be eventually exploited in the future if deemed necessary according to algorithm development and cal/val activity when HydroGNSS wil be in-orbit. In particular, soil texture can be used as input to the terrain dielectric constant model; porosity can be used to transform a saturation index between 0 and 1 into a volumetric soil moisture value in case of a change detection type of algorithm.

  Currently they are not used in the processor

- Soil small scale roughness (surface height standard deviation). It strongly affects the soil scattering properties, but unfortunately neither dynamic nor static reliable products exist.

  Currently it is used in the frame of the semi-empirical model (alfa release of the L2OP) to account for the reduction of the specular reflectivity due to the diffuse scattering from surface roughness. A static product is derived as described in the sequel from historical datasets, including:

  - o SMAP SSM, Vegetation Optical Depth, Vegetation Water Content and SSM at 36 km (or 9 km); 32MB at 36 km per day, 87MB at 9 km per day (https://podaac.jpl.nasa.gov/SMAP)
  - o Processing of historical GNSS-R data, such as CyGNSS (see previous sections in this document)

  Although possibly valuable, it is not used in the neural network algorithm since it is now limited to the tropical latitude range and an extension to higher latitudes requires a significant amount of HydroGNSS acquisitions.

Dynamic

- Vegetation Optical Depth (VOD) at a resolution of 6 km or better, or any other quantity that can act as a proxy of the VOD, such as Vegetation Water Content (WVC). A dynamic, almost real-time, dataset set is highly recommended. Depending on the availability of satellite products at the time of the HydroGNSS launch it can be provided by difference sources, among which:

  - o SMAP Vegetation Optical Depth and Vegetation Water Content, 32MB at 36 km per day, 87MB at 9 km per day (the baseline) (*https://podaac.jpl.nasa.gov/SMAP*)
  - o From MODIS Normalized Difference Vegetation Index (NDVI), 5 km resolution, every 16 days, 100MB (*https://modis.gsfc.nasa.gov/data/dataprod/mod13.php*), MOD13C1 product and Surface Reflectance products, 5 km resolution, daily, (https://modis.gsfc.nasa.gov/data/dataprod/mod09.php), MOD09CMG product to calculate Normalized Difference Water Index (NDWI)
  - o CCI Above Ground Biomass data updated yearly, 18GB full resolution (*https://climate.esa.int/en/odp/#/project/biomass*)

  Currently the processor uses SMAP as the primary source of VOD and VWC information. If SMAP data are not available or they do not have suitable quality the processor derives NDVI and NDWI from MOD09CMG MODIS data.

- Snow cover extension and soil freeze/thaw state. When the soil is frozen its dielectric constant and thus the reflectivity significantly decreases. In case of soil covered by snow the reflectivity can also change, especially when the snow is wet. Thus, the snow cover extension and freeze/thaw state data can be useful to filter out reflections where the reflectivity is not reliable for retrieving soil moisture, as specified later in this document.

Currently snow cover is added to the L1B product and exploited to filter out affected reflections. The information about freeze/thaw state is not guaranteed in the due time and then is not used at this stage.

3.4 OUTPUT DATA

The L2OP-SSM processor can generate both L2G and L3 products. The spatial resolution of both products and the temporal resolution of the L3 product (Nd days) can be selected through the configuration file. The spatial resolution can be set according to the standard cell size of the EASE Grid V2 projection, i.e., 25, 12.5 and 9 km. A detailed specification of the product netcdf format is contained in AD.1 and AD.5. Hereafter we provide a summary of the main quantity of interest for science and applications.

3.4.1 *L2G product content*

The product contains the retrieval of soil moisture along each individual track collected during a 6-hour time span, after averaging the L1B reflectivity within each EASE Grid cell. Data are formatted as vectors providing mean geographical coordinates and timing of the averaged reflectivity's, mean reflectivity's in Right and Left polarizations, number of reflections averaged in each cell that passed the quality checks (both in Right and Left polarization), the uncertainty in reflectivity estimates (standard deviation of the reflectivity's within each cell), 32 bits of quality flags, the processed signal (single polarization or dual polarization, GPS or Galileo), the predominant land cover class, the mean DEM elevation and slope, and the AGB derived from the static auxiliary data. In addition, a raster format soil moisture map is provided in EASE Grid V2 projection, cropped in the geographical area covered by the specular point track (adding an empty line on top and bottom and empty column at the left and right side) and containing *Not a Number* (NaN) in the cells not covered by the specular point track. Top- left pixel coordinates of the map are provided to make reference to the global EASE Grid reference system. An example of L2G trackwise product is shown in Figure 3-1, left panel. Blue and black circles are specular points that passed or not passed the quality checks, respectively. Blue cells are those where the mean reflectivity of good specular points where computed.

3.4.2 *L3 product content*

The content of the L3 product is substantially identical to that of L2G, although it is a unique file spanning the Nd days, eventually aggregating within each cell specular points coming from more than one track and from both HydroGNSS satellites. An example of L3 product in the same area of the L2G one is shown in Figure 3-1, right panel; it is understood that the L3 product will extent in principle over the entire globe after Nd days of data acquisition, but many cells will be empty if only two HydroGNSS satellites are in orbit.

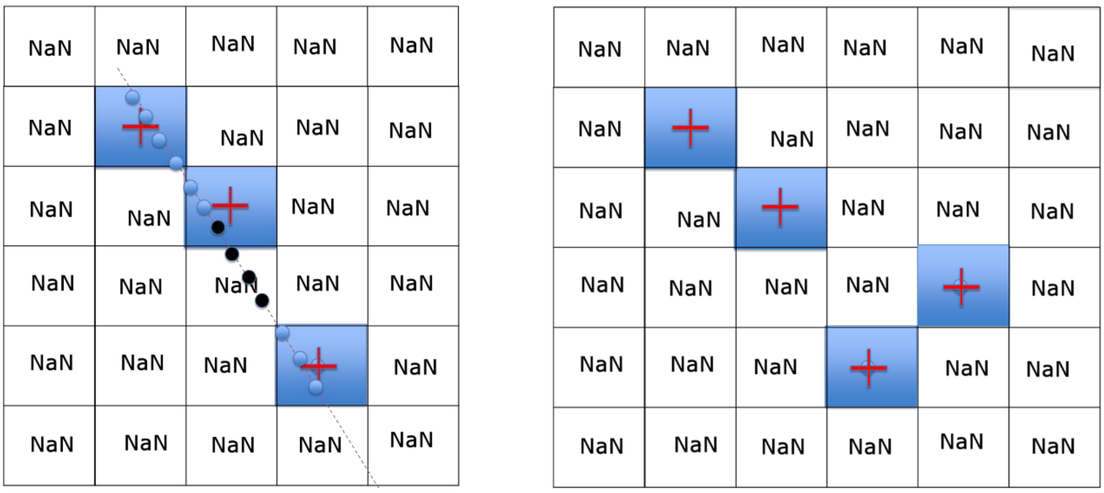


Figure 3-1. Example of L2G (left panel) and L3 (right panel) of soil moisture product in raster format.

3.5 DATA PRE-PROCESSING

The L2OP processor has two options, i.e., possibility to run the semiempirical algorithm or the neural network algorithm. The switch option is stored in the configuration file.

The data pre-processing software module injects all the needed data, quality check and combine them onto the same EASE Grid Global Cylindrical projection version 2.0. The size D of the EASE Grid cells, in km, is provided through the configuration file. A value of D=25 km is the baseline, but better resolutions (12.5 km and 9 km) are implemented and can be attempted in the future. For each EASE grid cell, the L1B data passing the quality checks and collected within a time frame (near instantaneous as for L2G, or within Nd days in case of L3 products) are averaged within the cell. Average is computed in the power unit (not dB). Note that different steps for quality checking and filtering out the input data are foreseen:

- Quality check and eventually filtering out individual specular points before averaging.
- Quality check and eventually filtering out the aggregate reflectivity within the cell (after averaging).
- Quality check of the dynamic auxiliary data.
- Quality check of the final soil moisture estimates

The integration time Nd for the L3 product can be selected through the configuration file. We note that with only one or two satellites not all the EASE grid cells can be covered in Nd days. Then the trade-off between spatial resolution (cell size D) and temporal resolution (Nd) will be made during Commissioning and future cal/val activity. Now, in the L2PP, alfa and beta releases of the processor a baseline of D=25 km is considered for L2G product. As for L3, the baseline is Nd=1 day of time resolution and D=25 km of spatial resolution. A maximum of Nd=3 days is suggested to match the temporal resolution of SMAP and SMOS.

The main pre-processing steps are:

- **Getting input** time frame, i.e., the first epoch and last epoch when the product shall be released.
- **Get configuration files** to set the processor modes, the input auxiliary files folders, and select the algorithm to be used, as well as its coefficients.
- The processor shall **identify the granules** to be processed. It is a generic six-hour block as for L2G product. It spans Nd days (the time resolution of the product) as for L3 product, including both satellites acquisitions.
- **Selection of HydroGNSS files** in the input time frame.
- Open the selected files and **read HydroGNSS** L1b product variables, among which: latitude and longitude of SPs, time of acquisition, DDM (the latter is optional as stated in the configuration file), reflectivity, NBRCS, quality flags, SNR, transmitter ID (e.g., PRN of GPS), degree of coherence (if available in L1B), static or dynamic data eventually available in the L1b product;
- **Single SP quality checks** and data filtering out. It consists in keeping L1b individual reflections that pass a number of quality checks; they are based on the quality flags included in the L1B product (provided the related field in L1B file is correctly filled) and internal checks of the processor. All the thresholds below are stored in the algorithm configuration files. The following checks are implemented, other can be envisaged in feature releases (below labelled as TBC):
    - Keep individual SP’s data with SNR greater than a threshold *Min_SNR* stored in the configuration file and common to all channels (2 dB in alfa and beta release, can be changed later on).
    - Keep individual SP’s with reflectivity higher than a threshold *Min_Reflectivity* common to all channel (-45 dB in alfa and beta release, TBD later on).
    - Keep individual SP’s where the snow cover map indicates “no snow”
    - Keep individual SP’s with antenna gain larger than a threshold *Min_antenna_gain* common to all channels.
    - Keep individual SP’s with incidence angle larger than a threshold *Max_incidence_angle,* common to all channels.
    - Keep individual SP’s with coherence index smaller or equal than a threshold *MaxSPCoherence* (higher values could correspond to the presence of a water body), common to all channels.
    - The following quality flags in the L1b metadata are used to filter out individual SP’s, i.e., *DirectSignalInDDM*, *Noise Kurtosis*, *the likely presence of RF interference, the likely presence of a correlator anomaly*, as reported in the ICD document.

  The following are TBC as some of them are not defined in detail yet, their usefulness or their availability must be verified.

    - *Coast proximity flag*

    - Keep individual SP's where the freeze/thaw product (from SMAP or HydroGNSS) indicates thawed conditions.
- **Gridding**. Identify to which EASE Grid cell (row/column coordinates) a data sample pertains and accumulate (add in linear units) the reflectivity, NBRCS and DDM (the latter is optional as stated in the configuration file) corresponding to the same cell that passed the quality checks. For each cell the number of accumulated samples shall be counted and stored. In case of L2G product, they pertain to an individual track. In case of L3 product they can pertain to different SP tracks, different HydroGNSS satellites and even different days if Nd>1. Divide the accumulated value by the number of accumulated sample (averaging).
- **Count the number of reflections** in a grill cell of both LHCP and RHCP good signals and derive the most probable value of the quality flags in the L1b product among the aggregated SP's.
- **Read static data** from the proper static auxiliary data folder and associate them to each non-zero cell of the EASE Grid: DTM parameters (mean height, mean slope, standard deviation of height), land cover (from CCI LCC map), small scale roughness (from AFS map), vegetation biomass (from CCI Biomass map), soil texture map (still TBD). Names of the static auxiliary files are in the conf file and will be reported in the output product among the global attributes of the output SM L2G/L3 product. The folders with the static data are also specified in the configuration file. In the beta version the static data are already stored in the same EASE Grid projection and same resolution of the HydroGNSS gridded data, so they can simply be stacked with the gridded reflectivity maps.
- **Search for the last available SMAP** product in the proper dynamic auxiliary data folder, as specified in the configuration file. The last three days of daily Level 3 SMAP product shall be selected at 36 km resolution (use of 9 km resolution SMAP data can be foreseen in the future).
- **Search for the** last available MODIS data in the proper dynamic auxiliary data folder (16-day MODIS NDVI and radiances to generate NDWI).
- **Check freshness of SMAP** data. If SMAP products are available, check whether available SMAP data are not earlier than *Oldest_SMAP* days before the acquisition time of the HydroGNSS data to be processed (*Oldest_SMAP* is in the configuration file, value of 16 days as a baseline).
- **If yes, pre-process SMAP** going to the following steps in the nested bullets.
    - Quality check SMAP data, retaining only data that pass the "*recommended quality flag*" check in the SMAP product. Merge the SMAP data (i.e., variables VOD and VWC from L3 product) of three days; this is needed to have all the EASE Grid cells filled by SMAP
    - Resample or aggregate SMAP data (i.e., VOD and VWC) over the EASE Grid reference cells at D km resolution. In case of 36 km resolution SMAP products they shall be geographically resampled over the D km resolution grid of the reflectivity. In case of 9 km resolution SMAP, they shall be aggregated over the same EASE Grid (avoid just undersampling).
    - **If SMAP in a cell is not good** go to step **Check freshness of MODIS**
    - **Otherwise select the algorithm that uses SMAP**
    - Go to **Apply selected retrieval algorithm.**
- **If not, go to the following step** checking for MODIS data
- **Check freshness of MODIS.** Check whether available MODIS data are not earlier than *Oldest_MODIS* days before the acquisition time of the HydroGNSS data to be processed (*Oldest_MODIS* in the configuration file, value of 20 days as a baseline).
- **If yes, set algorithm that uses MODIS** data and go to the following steps in the nested bullets
    - Quality check and filter out MODIS data based on MODIS quality flag (TBD), and suitable range of NDVI (TBD)
    - Compute NDWI from MODIS Surface Reflectance products. At least 10 days of MODIS Surface Reflectance product should be used in order to overcome the cloud cover problem and retain the cloud free pixels to generate a NDWI map as wide as possible. Cloudy pixels are filtered out from band 2 (NIR) and band 5 (SWIR) of the MODIS reflectance by using the band "Coarse Resolution Internal CM" as an internal cloud mask as descripted in MODIS "MOD09_UserGuide_v1.4" document. Then NDWI is calculated by NDWI = (NIR – SWIR)/(NIR + SWIR) formula.

   - Resample or aggregate MODIS data (i.e., NDVI and ND) over the EASE Grid reference cells at D km (see case of SMAP).
   - **If MODIS in a cell is not good** go to step **select algorithm without dynamic auxiliary**
   - **Otherwise select the algorithm that uses MODIS.**
   - Go to **Apply selected retrieval algorithm.**
- **If not** (i.e., neither SMAP data nor MODIS data are available in the cell), **select the algorithm without dynamic auxiliary**, i.e., the processor shall generate SM product with a non-optimal algorithm and rice a flag "*no_dynamic_auxiliary*" to indicate this. Then go to **apply selected retrieval algorithm**
- **Apply selected retrieval algorithm** to the stacked data considering two options
   - **Apply dual polarization algorithm** in case the number of aggregate good reflections in RHCP is larger than *Min_RHCP_SPs*
   - **Apply single polarization algorithm** in case the number of aggregated good reflections in RHCP is smaller than *Min_RHCP_SPs*

3.6 Semi-empirical forward model

Considering both a coherent and incoherent contribution coming from the surface the total received signal power is $P_r^S = P_i^S + P_c^S$ (see section 2.3). The incoherent diffuse scattering can be due to soil roughness and to vegetation volume scattering, then the effective reflectivity measured by the system is modelled as three contributions (Yueh et al., 2022):

$$\Gamma_{eff}(\theta) = \frac{(P_c^S+P_i^S)(4\pi)^2(R_r+R_t)^2}{\lambda^2 G^t\, G^r\, P_t} = \frac{P_c^S(4\pi)^2(R_r+R_t)^2}{\lambda^2 G^t\, G^r\, P_t} + \sigma_{soil}^o \frac{(R_r+R_t)^2 A_S}{4\pi R_r^2 R_t^2} + \sigma_{veg}^o \frac{(R_r+R_t)^2 A_S}{4\pi R_r^2 R_t^2} \quad (3.1)$$

In the framework of the Kirchhoff approximation of surface scattering the first term (the near-specular, generally coherent one) is related to the Fresnel reflectivity (square module of the Fresnel reflection coefficient) diminished by a term that accounts both for the radiation scattered in other directions due to small-scale roughness and the attenuation of the vegetation. The second term (a diffuse incoherent one) is also proportional to the square module of the Fresnel reflection coefficient multiplied by a factor that is dependent on surface roughness (including large-scale). The third term is the vegetation incoherent volume scattering. Then the following applies:

$$\Gamma_{eff}(\theta) = e^{-2b\cdot VWC\cdot \mathbf{sec}(\theta)}|R(SM)|^2\left[e^{-4k^2h^2cos^2(\theta)} + F(h)S\right] + \sigma_0^o\left(1 - e^{-2b\cdot VWC\cdot \sec(\theta)}\right)S \quad (3.2)$$

Where θ is the incidence angle at the specular point, VWC is the Vegetation Water Content, R is the Fresnel reflection coefficient, which is the parameter function of soil moisture, SM is the soil moisture, k is the wave number (L-band), h is the surface small-scale roughness standard deviation, σ is the BRCS of an infinite vegetated layer, and $F(h)$ is the roughness dependent factor of the NBRCS. The vegetation optical depth is expressed by an empirical model as $\tau = 2b \cdot VWC \cdot \sec(\theta)$, where coefficient b is generally dependent on the vegetation type and growing stage. In eq. 3.2, S is a geometric scaling factor $S = (R_t + R_s)^2\, A_s\, /4\pi R_r^2\, R_t^2$ that can be derived from eq. (2.1) and eq. (2.2). Referring to eq. (2.3) the same factor can be more accurately written as:

$$A_S = \int \chi^2(\tau, f_D)dA \quad (3.3)$$

The Fresnel reflection coefficient for RHCP incident and LHCP observed polarizations (R_{LR}), and for RHCP incident and RHCP scattered (R_{RR}) are related to the Fresnel coefficient for horizontal and vertical linear polarizations, R_H and R_V. The relevant formulas for an incidence angle θ are:

$$R_{LR} = \frac{1}{2}(R_{VV} - R_{HH});\ R_{RR} = \frac{1}{2}(R_{VV} + R_{HH});\ R_H = \frac{\sqrt{\epsilon_{soil} - sin^2(\theta)} - \cos(\theta)}{\sqrt{\epsilon_{soil} - sin^2(\theta)} + \cos(\theta)};\ R_V = \frac{\epsilon_{soil}\cos(\theta) - \sqrt{\epsilon_{soil} - sin^2(\theta)}}{\epsilon_{soil}\cos(\theta) + \sqrt{\epsilon_{soil} - sin^2(\theta)}} \quad (3.4)$$

The model for the soil dielectric constant ε_{soil}, which is recognized more accurate and used both in SMAP and SMOS ground processing, is the so-called Mironov model described in (Mironov et al., 2009).

The term within square parenthesis in eq. (3.2) is the most challenging to be determined in order to invert the forward model and retrieve the soil moisture dependent Fresnel reflectivity. It is due to both small-scale roughness reduction of the coherent specular reflection and to the incoherent diffuse scattering of the intermediate and large-scale roughness. In the absence of a clear distinction between the two terms it can be grouped in a unique "bulk" surface scattering parameter (Attenuated Forward Scattering: AFS) as in the following:

$$\Gamma_{eff}(\theta) = e^{-2\tau} \cdot AFS \cdot |R(SM)|^2 + \sigma_0^o(1 - e^{-2\tau})S \tag{3.5}$$

Note that, besides the target soil moisture, other unknown parameters appear in eq. (3.5), namely AFS, τ and σ_0^o, whereas S is only a geometric factor. Among them, an assumption that can be made is considering AFS and, at some extent, σ_0^o as static time invariant parameters. The vegetation opacity, very much associated to the water content, is instead likely a quantity that varies over time.

The unknown quantities were estimated before launch from an experimental set of CyGNSS and SMAP data as explained in section 3.7.2. Of course, this could be done only in the latitude range covered by CyGNSS, so that the work should be repeated during commissioning using at least one month of HydroGNSS data.

3.7 RETRIEVAL ALGORITHM

3.7.1 *Multilinear regression algorithm (L2PP placeholder)*

Note that, in case the diffuse contribution from soil roughness and vegetation volume scattering can be neglected, by considering the effective reflectivity in decibel units, the following applies:

$$\Gamma_{dB}(\theta) = 20log_{10}|R(SM)| + A(\theta) \cdot VWC + B(\theta) \cdot h^2 = |R(SM)|_{dB} + A(\theta) \cdot VWC + B(\theta) \cdot h^2 \tag{3.6}$$

The equation becomes linear, i.e., the reflectivity in dB is linearly related to the module of the Fresnel reflection coefficient (expressed in dB), the VWC and the roughness variance (h^2). The coefficients of the linear relation are function of the incidence angle.

$$A(\theta) = -2bsec(\theta) \qquad B(\theta) = -4k^2cos^2(\theta) \tag{3.7}$$

This result justifies the method proposed in (Clarizia et al., 2019), where a multivariate linear regression was calibrated using a collocated dataset of CyGNSS reflectivity, SMAP vegetation opacity, SM and roughness. In fact, by assuming a simple linear relation between the module of the Fresnel reflection coefficient and SM, and replacing the roughness variance h^2 with the roughness coefficient $\sigma=4k^2h^2$ introduced in the SMAP processing and available through the SMAP L3 products, one can reasonably write:

$$SM = a_j \cdot \Gamma_{dB}(\theta) + b_j \cdot \tau' + c_j \cdot \sigma' + d_j \tag{3.8}$$

According to (Clarizia et al., 2019) the prime superscript indicates standardized variables; it means that the dynamic range of the dependent variables, i.e., vegetation opacity τ and roughness standard deviation h, are kept in the interval [0,1]. The standardization of the independent variables is done by the following.

$$\tau' = \frac{\tau - \tau_{min}}{\tau_{max} - \tau_{min}} \qquad \sigma' = \frac{\sigma - \sigma_{min}}{\sigma_{max} - \sigma_{min}} \tag{3.9}$$

Subscripts *max* and *min* indicate maximum and minimum values of the variables. The subscript j in the regression coefficient a, b, c, d indicates the possibility to stratify the training sample to do the regression and retrieve different coefficients for different zones of the Earth. One can retrieve different coefficients for different sites, for different land use / land cover categories or even for each EASE grid cell.

The coefficients of the retrieval formula in eq. (3.8) can be tuned using a suitable training set. In case the vegetation optical depth τ is not available, it can be replaced by a proxy that in case of forest can be the biomass and in case of crop the vegetation height. Assuming index i represents each sample of the training set, for each land use / land cover category j, the coefficients a_j, b_j, c_j, d_j can be derived as the ones that minimize the following object function:

$$\sum\left[SM_i - a_j \cdot \Gamma_{dB,i}(\theta) - b_j \cdot \tau_i' - c_j \cdot \sigma_i' - d_j\right]^2 = \text{min} \tag{3.10}$$

The value of the multivariate regression coefficient and the standardization quantities in (Clarizia et al., 2019) regards the case of CyGNSS product version 2.0 and reflectivity computed using eq. (2.1). For the global case (i.e., a unique set of coefficients for all the world) the coefficients are reported in Table 3-2. It is worth reminding that in (Clarizia et al., 2019) the authors applied a "corrective" bias after the regression, which was needed to compensate for any non-linearity in the real data with respect to what is assumed in this approach. Another option to account for non-linearity consists in adding powers of the input quantities in the regression, i.e., considering a polynomial regression. Anyhow, more sophisticated algorithms were considered for alfa- and beta-release of the L2OP-SSM processor exploiting the full model in eq. (3.1) or an ANN approach. Then, it was not foreseen to develop the placeholder further as it was used mainly to assess the E2ES software chain.

As for the performance assessment task based on the placeholder, we have used a set of E2ES outputs to train the algorithm. H-SAVERS was run for a range of input geophysical variables extracted randomly within reasonable intervals (e.g., soil moisture, forest biomass, crop height, soil roughness, etc.) in the sites identified to assess soil moisture retrieval performances (San Luis Valley).

Table 3-2. Coefficients of the global algorithm using CyGNSS reflectivity and SMAP vegetation opacity and roughness parameters (RF#4).

		a	*b*	*c*	*d*
CyGNSS v2.0	global	0.88	0.54	-0.18	-0.43

3.7.1.1 *Actual implementation in the E2ES*

Not relevant for the PDGS algorithms. See AD.8 for details.

3.7.2 *Semi-empirical model fitting of stratified samples (L2OP-SSM alfa release)*

The L2OP alfa-release retrieval algorithm is based on the inversion of the semi-empirical model depicted in section 3.6. This was considered a suitable approach in the first stage of the mission development to perform E2E simulations keeping the control of the physical mechanisms of the target interaction and better understanding the simulation results as a function of system parameter errors and auxiliary data errors. This algorithm can be also selected in the PDGS during operation to test its performances. With respect to the 'placeholder' the following improvements were implemented:

- The inversion algorithm minimizes a cost function based on a physically based forward model and thus is able to account for the non-linearity of the relation between observables and target surface parameters inherent in the physical forward model.
- A suitable stratification is carried out by tuning different sets of algorithm coefficients for different strata. Stratification is based on a geographical segmentation of the EASE grid cells combined with the land use / land cover class of the cells. The number of cells and the aggregation of class of the map can be optimized as the algorithm development activity will progress. CyGNSS data at global scale were analysed for setting the internal configuration file containing the geographical stratification map and the algorithm coefficients.
- A static roughness map can be estimated by a time series of GNSS-R data combined with SMAP data providing the reference soil moisture. The timeseries should be long enough to provide a full coverage of the EASE Grid map at the prescribed resolution. A long time series (typically 3-6 months) would improve the quality of the roughness map. Before launch, it has been derived by combining a longer series of SMAP and CyGNSS data, although this is limited to the tropical latitude band.
- The right-right polarization new observable is exploited in the L2OP alfa release in a simple way based on a purely empirical approach as the forward model for RHCP is not established yet in the scientific community and needs more research. The inclusion of RHCP was only possible in the frame of the E2ES processing chain, as no real RHCP measurements are currently available for training.

In order to estimate AFS and σ_0^o in eq. 3.5, it is needed a collocated (in space and time) dataset of measured (equivalent) reflectivity Γ_{eff}, soil moisture SM (from which it is possible to derive the Fresnel reflection coefficient R) and vegetation opacity τ. Currently this was done by relying on CyGNSS L1b products providing reflectivity, and collocated SMAP products providing SM and τ. A suitable timeframe covering different terrain geophysical conditions was considered. Then the following objective function was minimized in the mean square sense to retrieve AFS in each EASE Grid cell (the AFS map) and σ_0^o in each stratum (a σ_0^o table):

$$f(x) = \Gamma_{eff}^{CYG} - [e^{-2\tau} \cdot AFS \cdot |R(SM)|^2 + \sigma_0^o(1 - e^{-2\tau})S] \tag{3.11}$$

$R(SM)$ is computed using the Fresnel reflection coefficient formula and the model of soil permittivity as a function of SM (a soil texture map could be considered when available). S depends on known geometric quantities.

As soon as the two unknown static parameters AFS and σ_0^o are known, it is straightforward writing, starting from eq. (3.5):

$$10 \cdot log10[\Gamma_{eff}(\theta) - \sigma_0^o(1 - e^{-2\tau})S] = -10 \cdot 2\tau \cdot log10(e) + 10 \cdot log10(AFS) + 10 \cdot log10(|R(SM)|^2)$$

From which we can write in decibel units:

$$|R(SM)|_{dB}^2 = \left(\Gamma_{eff}(\theta) - \sigma_0^o \cdot S \cdot (1 - e^{-2\tau})\right)_{dB} + 20 \cdot \tau \cdot log_{10}e - AFS_{dB} \quad (3.12)$$

Eq. (3.12) is a linear relationship between the Fresnel reflection coefficient (in dB) and the AFS (in dB), the vegetation optical depth and the effective measured reflectivity subtracted by the vegetation volume scattering (if significant). Considering that some uncertainties and weakness of the forward model may be still present and must be compensated for and considering the presence of instrumental noise and calibration errors, we can empirically generalize the above relationship, admitting different weights of the terms and eventually a polynomial relationship. Thus, a minimum error variance regression algorithm can be implemented to estimate the Fresnel reflection coefficient assuming the knowledge of the two time-invariant parameters (AFS and σ_0^o) and of the vegetation optical depth available from other sources. The coefficients of the regression are to be retrieved from a training set independent on that used to estimate AFS and σ_0^o, by minimizing the following object function in the mean square sense:

$$f(x) = \left\{|R(SM)|_{dB}^2 - \left[a \cdot \left(\Gamma_{eff}^{CYG} - \sigma_0^o \cdot S \cdot \left(1 - e^{-2\tau}\right)\right)_{dB} + b \cdot \tau + c \cdot AFS_{dB} + d_i \cdot \beta_i + e\right]\right\}^2 \quad (3.13)$$

Currently, the training of the algorithm was carried out again by a collocated dataset of CyGNSS and SMAP products collected in a different time frame with respect to that used for estimating AFS and σ_0^o.

The new parameters β_i introduced empirically in the minimization, together with the related unknown multiplicative coefficients d_i, can be the new co-polarized observable from HydroGNSS (i.e., the Right-Right polarization), the degree of coherence, the waveform trailing edge or a combination of them. This will be assessed during further development and testing of the algorithm.

Once the Fresnel reflection coefficient is retrieved from a new measurement of the effective reflectivity, the inversion of the Mironov formula (Mironov et al., 2009) to retrieve the soil moisture is straightforward. A global texture map (e.g., map of clay fraction) could be envisaged to implement such inversion in a later release to work at global scale.

The alfa release of the processor assumes a simplified version of the approach, which consists in assuming a linear relation between the Fresnel reflection coefficient in decibel and the soil moisture. The use of a non-linear model did not provide significant improvements in the performances. In such a case the algorithm reduces again to a simple multivariate regression with a modified input vector since we must minimize the following object function:

$$f'(x) = \left\{SM - \left[a' \cdot \left(\Gamma_{eff}^{CYG} - \sigma_0^o \cdot S \cdot \left(1 - e^{-2\tau}\right)\right)_{dB} + b' \cdot \tau + c' \cdot AFS_{dB} + d'_i \cdot \beta_i + e'\right]\right\}^2 \quad (3.14)$$

The algorithm can be considered semi-empirical, as it accounts for a functional relationship between observed reflectivity and bio-geophysical parameters derived from physical considerations, but also enables introducing additional auxiliary information or different weights of the predictive quantities that are deemed useful for the inversion.

3.7.2.1 **Novel HydroGNSS observables**

The new observables from HydroGNSS were introduced in the L2OP processors (alfa and beta releases) in a purely empirical way in the frame of the semi-empirical approach as well as the neural network approach (see section 3.7.3). Then, they will be possibly included in the physically based framework whenever new models will be made available.

As for soil moisture the availability of the co-polarized DDM is deemed extremely promising, even if the low magnitude of RR reflections could result in a poor signal to noise ratio (SNR) in some areas, particularly dense forests. Regarding the “complex channel”, i.e., the complex signal from the nominal specular point, it is not expected to provide a significant contribution over targets where the signal is expected predominantly incoherent (i.e., agricultural areas, natural and vegetated terrain with variable topography of interest for soil moisture applications). Nevertheless, a by-product of that observable, i.e., the estimate of the degree of coherence, can be valuable to be included in the retrieval as it may help predicting the DDM shape and the reflectivity according to the amount of coherent and incoherent signal.

Regarding the dual polarization capability of HydroGNSS, it is known that the roughness represents a common multiplicative factor affecting the reflectivity in both polarization and in fact it has been observed both empirically and theoretically that the linear ratio between RR/LR tends to remove the effect of the roughness and highlight the effect of soil moisture. This is shown for instance in Figure 3-2, where the H-SAVERS reflectivity simulations along a specular point track are shown for both single polarizations and their ratio. Note that the LR and RR are very much affected by the topography, whereas the linear ratio RR/LR exhibits a trend that is inversely related to the soil moisture.

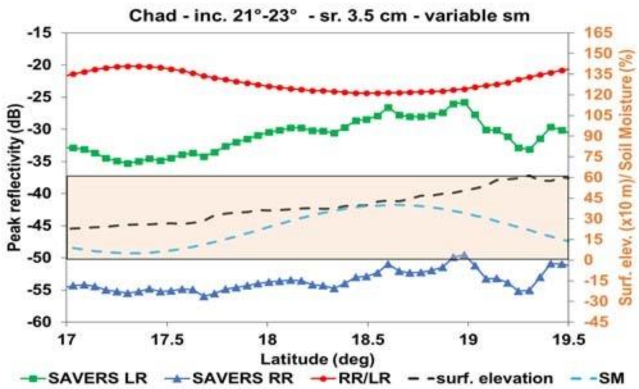


Figure 3-2. LR, RR and RR/LR bare soils simulations along a track passing over a gentle topography with variable soil moisture. The topography is reported with the black dashed line and the soil moisture with cyan dashed line in the inset rectangle.

The RR observables and the degree of coherence can be exploited within the semi-empirical approach depicted by eq. (3.12) considering that a linear ratio corresponds to a difference in dB by training the following multivariate linear regression:

$$f(x) = |R(SM)|^2_{dB} - \Big[a \cdot \left(\Gamma^{CYG}_{LR,eff} - \sigma^o_{LR,0} \cdot S \cdot \left(1 - e^{-2\tau}\right)\right)_{dB} + b \cdot \tau + c \cdot AFS_{dB} + \\ + d \cdot \left(\Gamma^{CYG}_{RR,eff} - \sigma^o_{RR,0} \cdot S \cdot \left(1 - e^{-2\tau}\right)\right)_{dB} + e\Big] \qquad (3.15)$$

The vegetation parameters σ^o_0 must be estimated separately for LR and RR polarizations, whereas the AFS parameter can be in common. Refer to section 4 for the estimation of the parameters for each grid cell and strata. It is worth noticing that $\sigma^o_{LR,0}$ was estimated at global scale from CyGNSS data in the tropical latitude range, whereas $\sigma^o_{RR,0}$ initially is set to zero and be estimated only when a suitable amount of co-polarised measurements will be available from HydroGNSS.

3.7.2.2 **Actual implementation in the E2ES**

Not relevant for the PDGS algorithms. See AD.8 for details.

3.7.3 *Artificial Neural Network (L2OP-SSM beta release)*

The beta release of the L2OP-SSM processor includes a data driven Artificial Neural Network (ANN) approach placed next to the physically based approach in the processor. As explained in section 3.1, the final approach that will be exploited in the PDGS during HydroGNSS operations will be chosen between the semi-empirical model and the ANN one as a result of the algorithm development and Cal/Val activities before launch and eventually later during commissioning.

The ANN algorithm is a feed-forward multilayer network built with MATLAB©'s Deep Learning Toolbox. The back-propagation rule serves as the foundation for the network training phase. The definition of the network structure was based on a trial-and-test procedure where, beginning with a standard condition, the number of hidden layers and neurons was iteratively raised, training and testing the network itself, until a suboptimal condition was obtained. The available activation functions also went through this trial-and-test process.

As a result of the suboptimization process, the structure of the ANN implemented in the processor is a network with 3 hidden layers, each containing 14 neurons, and a hyperbolic tangent sigmoidal activation function, as depicted in Figure 3-1. The network structure that results from this procedure is capable of preventing the overfiting issue.

The training set of target soil moisture was derived from the Soil Moisture Active Passive (SMAP) mission Level 3 (daily ascending and descending) products collocated in time and space with CyGNSS gridded data. A 25 km grid sampling was considered. The network was then applied to an independent set of CyGNSS data gridded at the same resolution to generate and assess the product accuracy in comparison to SMAP moisture retrievals. The SM product generated at 25 km resolution was also tested against fully independent in-situ measurements available through the In-situ Soil Moisture Network (ISMN). The network was fed not only by CyGNSS data (different observables and ancillary information, such as the incidence angle), but also by static auxiliary data gathered from other sources, e.g., the surface topography, and dynamic bio-geophysical parameters from other satellites, e.g., Vegetation Optical Depth from SMAP.

The dataset was divided into two parts, 10% for training and 90% for testing (i.e., comparing with data different from those used in the training phase), to achieve the statistical independence required to ensure certain generalization capabilities of the proposed algorithm. Moreover, during the network training phase, the training dataset was then randomly divided again into three parts, respectively equal to 70%-15%-15% of the data, according to the so-called "Early Stopping Rule".

Note that according to the quality checks during data preprocessing, not all inputs are available or have the appropriate quality to be used for retrieval. For instance, the SNR quality check could reject high noisy reflections in right polarization, or the SMAP and MODIS data could be not available or not of the appropriate quality. Therefore, different ANN models are considered according to the available input data. In the current implementation of the processor the inputs to the network are summarized in

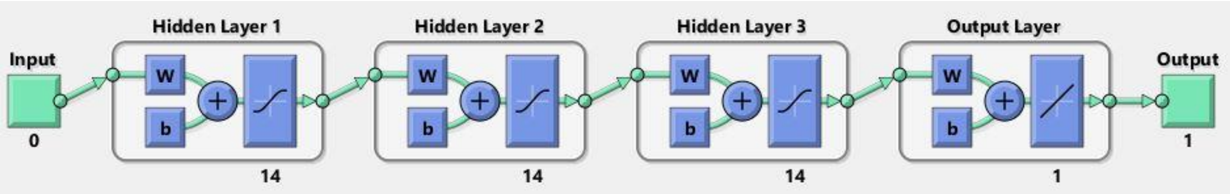


Figure 3-2. Architecture of Neural Network

Table 3-3 as colored boxes for the different cases of input availability. Empty boxes denote input not used or eventually to be considered after Commissioning in the frame of future algorithm development activity.

Regarding the combination of GNSS signals that can be exploited, the following cases are foreseen:

1. L1 left: it processes only L1 using Left polarization
2. L1 dual: this corresponds to the 'dual polarization' signal, i.e., it processes only L1 using both Left and Right (if Right passes the quality checks)
3. E1 left: it processes only E1 using Left polarization
4. E1 dual: it processes only E1 using both Left and Right polarizations (if Right passes the checks)
5. L1&L5 dual: it corresponds to the 3-channels mode, which processes L1/L5 left and L1 right. According to the outcomes of the quality flags, it may use four models working with L1 left (single polarization), L1 left/right (dual polarization), L1/L5 left (dual frequency), L1 left/right and L5 left (3-channels). The models are invocated depending on whether or not the L1 left/right or L5 left passes the quality checks.

6. E1&E5 dual: it corresponds to the 3-channels mode, which processes E1/E5 left and E1 right. According to the outcomes of the quality flags, it may use four models working with E1 left (single polarization), E1 left/right (dual polarization), E1/E5 left (dual frequency), E1 left/right and E5 left (3-channels). The models are invocated depending on whether or not the E1 left/right or E5 left passes the quality checks.

It must be pointed out that before the launch only the ANN model of case 1 was reliably trained using satellite data (i.e., CyGNSS), whereas the other models could only be trained using simulated data over small sites.

3.7.3.1 Actual implementation in the E2ES

Not relevant for the PDGS algorithms. See AD.8 for details.

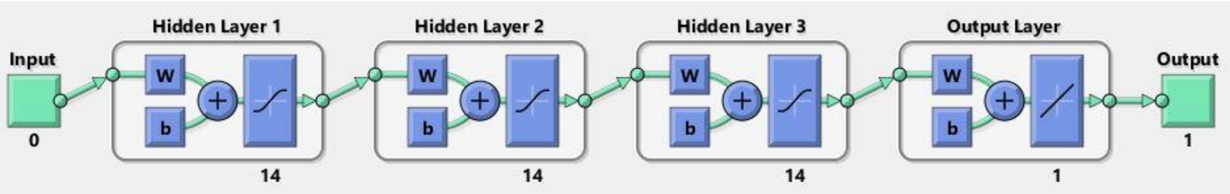


Figure 3-2. Architecture of Neural Network

Table 3-3. Current input data used by the ANN processor for different cases of availability of good inouts form HydroGNSS and auxiliary sources

	Good L1/E1 Left			Good L1/E1 Left and right			Good L1/E1 Left and L5/E5 Left			Good L1/E1 Left and Right and L5/E5 Left		
	SMAP available	MODIS (No SMAP)	No SMAP, No MODIS	SMAP available	MODIS, (No SMAP)	No SMAP, No MODIS	SMAP available	MODIS (No SMAP)	No SMAP, No MODIS	SMAP available	MODIS (No SMAP)	No SMAP, No MODIS
Reflectivity L1/E1 Left												
Reflectivity L1/E1 Right												
Reflectivity L5/E5 Left												
Reflectivity L5/E5 Right												
Incidence angle												
SNR L1/E1 Left												
SNR L1/E1 Right												
SNR L5/E5 Left												
SNR L5/E5 Right												
SMAP VOD												
SMAP WVC												
MODIS NDVI												
MODIS NDWI												
DEM height												

DEM height st. deviation												
DEM slope												
DEM slope st. deviation												
AGB												
Land use / land cover												
AFS												
Snow cover												
SNR L1/E1/L5/E5 Right/Left												
Reflect. L1/E1/L5/E5, R/L												
Antenna gain												
Antenna gain												
Incidence angle												
Coherence indec												
Freeze thaw												
L1B flags												
Wavef. Trailing Edge Lenght												
Coherency index												

	HydroGNSS observables		Static auxiliary data
	Dynamic auxiliary data		SP filtering

3.7.4 *Quality check on the final product*

A final check is applied to the retrieved soil moisture. Four thresholds are identified and stored in the configuration file: *SSMmin1*, *SSMmin2*, *SSMmax1*, *SSMmax2*.

If estimated SSM is SSM=<*SSMmin1* the retrieval is discarded (a Nan is posted). If *SSMmin1*>SSM=>*SSMmin2*, SSM will be set to SSM=*SSMmin2* and delivered setting a *flag_min* up. If *SSMmax2*>SM=>*SSMmax1*, SSM will be set to SM=*SSMmax1* and delivered setting a *flag_max* up. If SSM=>*SSMmax2* the retrieval will be discarded. The thresholds are in the algorithm configuration file. Note that *SSMmin2*=0, whereas the other baseline values for the threshold are: *SSMmax1*=0.5, *SSMmax2*=0.60, *SSMmin1*= -0.1

3.8 QUALITY FLAGS

Quality flags in the input L1B product will be reported in the L2G and L3 product considering the spatial and temporal aggregation:

- <*QualityFlag*>The most probable value of the L1B quality flags (modal value) among the observations averaged to derive the track-wise gridded (L2G) or gridded (L3) product will be associated to each soil moisture value.

Quality indices generated by the L2OP processor are specified in what follows. Threshold values used to set the quality flags are stored in the configuration file.

- <*NumIntegratedSPs_LHCP*>: Number of L1/E1 LHCP observations averaged to get a single soil moisture retrieval (the larger the number of observations the more reliable is the mean reflectivity estimate)
- <*NumIntegratedSPs_RHCP*>: Number of L1/E1 RHCP observations averaged to get a single soil moisture retrieval in case of the dual pol and three channels algorithm (the larger the number of observations the more reliable is the mean reflectivity estimate)
- <*NumIntegratedSPs_LHCP5*>: Number of L5/E5 LHCP observations averaged to get a single soil moisture retrieval in case of the dual frequency and three channels algorithm (the larger the number of observations the more reliable is the mean reflectivity estimate)
- <*Uncertainty*>*:* radiometric resolution of L1B LHCP observable computed as 10*log10(1+std/mean), where std and mean are the standard deviation and mean values of the reflectivity in the linear unit of the specular points averaged within the grid cell (smaller standard deviation denotes a smother response of the EASE Grid cell and a better retrieval of the reflectivity)
- <*SSMQuality*>: The quality flags are stored in a 32 bit word, each bit indicating bad (flag on, value is 1) or good (flag off, value is 0) quality with respect to a certain criterion. The following flags will be set "on" in each cell of the EASE grid in the following conditions:

  Bit 1. At least one of the flag below is on (it means the overall quality of the retrieval is not optimal)
  Bit 2. Standard deviation of L1/E1 LHCP reflectivity of averaged observations is greater than a threshold *MaxStdLHCP* (observations are affected by strong fluctuations)
  Bit 3. Standard deviation of L1/E1 RHCP reflectivity of averaged observations in case of the dual pol and three channels algorithm is greater than a threshold *MaxStdRHCP* (observations are affected by strong fluctuations)
  Bit 4. Standard deviation of L5/E5 LHCP reflectivity of averaged observations in case of the dual frequency and three channels algorithm is greater than a threshold *MaxStdLHCP5* (observations are affected by strong fluctuations)
  Bit 5. Mean SNR of L1/E1 LHCP observations within the cell from L1B product is lower than a threshold *LowSNR_LHCP*
  Bit 6. Mean SNR of L1/E1 RHCP observations within the cell from L1B product is lower than a threshold *LowSNR_RHCP*
  Bit 7. Mean SNR of L5/E6 LHCP observations within the cell from L1B product is lower than a threshold *LowSNR_LHCP5*

Bit 8. Mean surface height in the cell is greater than a threshold *HighDEMelevation* (soil moisture retrieval is challenging in mountainous areas)
Bit 9. Mean surface slope in the cell greater than a threshold *HighDEMslope* (soil moisture retrieval is challenging in mountainous areas)
Bit 10. Position of the absolute peak of the DDM and position of the geobounded peak on the Earth is larger than *Max_location_uncertainty* in km (high geolocation uncertainty)
Bit 11. Retrieved soil moisture was less than SSMmin2 and was set to SSMmin2
Bit 12. Retrieved soil moisture was larger than SSMmax1 and was set to SSMmax1
Bit 13. Retrieval was using single polarization algorithm
Bit 14. Retrieval was using dual polarization algorithm, single frequency
Bit 15. Retrieval was using dual frequency algorithms, single polarization
Bit 16. Retrieval was using MODIS dynamic auxiliary data
Bit 17. Retrieval was not using any dynamic auxiliary data
Bit 18. The mean coherence index in the cell is larger than a threshold MaxCellCoherence (there is a possibility the reflection is affected by the presence of water bodies)
Bit 19. TBD
Bit 20. TBD
Bit 21. TBD
Bit 22. TBD
Bit 23. TBD
Bit 24. TBD
Bit 25. TBD
Bit 26. TBD
Bit 27. TBD
Bit 28. TBD
Bit 29. TBD
Bit 30. TBD
Bit 31. TBD
Bit 32. TBD

3.9 Summary of processors features

Table 3-4. Mapping the algorithms and the data exploitation implemented in the different Level 2 SSM processors in case of using the processor within the H-E2ES for performance assessment and within the PDGS for operational product generation.

Algorithms and data	Use	placeholder	Alfa release	Beta release
Regression	E2ES PDGS	Yes No	Yes No	No No
Physically based	E2ES PDGS	No No	Yes Yes	Yes Yes
Neural network	E2ES PDGS	No No	No No	Yes Yes
Input checks		Basic	Basic	Yes
Output QI's		No	Basic	Yes
Static Auxiliary		DEM, CCI, AGB	DEM, CCI, AGB	DEM, CCI, AGB
Dynamic auxiliary	E2ES PDGS	Reference file No	Reference file SMAP	Reference file SMAP, MODIS
Other auxiliary	E2ES PDGS	No No	No No	No Snow, F/T, texture to be added
Dual pol	E2ES PDGS	No No	Yes No	Yes Yes
Coherence index	E2ES PDGS	No No	No No	No Yes

Spatial resolution	E2ES PDGS	25 km No	25 km 25 km	25 km 25, 12.5, 9 km
Temporal resolution	E2ES PDGS	Trackwise (L2G)	L2G L2G	L2G L2G, and L3 (Nd=1-3)

Table 3-4 summarises the functionalities included into the different processors released at this stage, as well as their capability to exploit input auxiliary data. This is specified, when needed, both when the processor is used in the frame of the H-E2ES for performance assessment purposes and in case it is integrated within the PDGS for operational processing and product generation.

4 CALIBRATION AND VALIDATION

The verification of compliance with software requirements and specifications, including the correct implementation of the algorithms described in this document, is straightforward according to the V&V plan. As for validation of the capability of the final product to meet the user requirements, this is a standard process during a satellite EO mission development and require availability of reference data about the quantity to be delivered (e.g., in situ data collected during the satellite overpass) to be compared to the satellite retrievals. We foresee the following steps and needs.

4.1 VALIDATION BEFORE LAUNCH

In the frame of algorithm development, data from GNSS-R airborne or satellite instruments are compared to soil moisture retrieved from other missions and in situ data. An airborne campaign would be important to acquire the new observations foreseen by HydroGNSS, i.e., dual polarization and coherent channel. The envisaged data are therefore:

- CyGNSS constellation L1B product as source of “standard” GNSS-R data
- Reference SSM retrievals from SMAP mission
- Reference SSM in situ data from the International Soil Moisture Network (freely available)
- Airborne GNSS-R campaign acquiring dual polarized reflections over a well-controlled site equipped with in SSM situ probes for reference purposes (when available).

The use of the datasets listed in the first three bullets represents the baseline for Cal/Val activities that are being carried out before launch and will be done later on. The data in the fourth bullet, although highly recommended, are not planned until now due to budget restrictions.

4.1.1 Semi-empirical model

Whilst for performance assessment scope the algorithm training was based on the output of the H-SAVERS simulator in the considered sites, the calibration of the algorithm at global scale was based on global data taken from CyGNSS and proper auxiliary information. A stratification of the whole sample set was also carefully designed. This task is being prosecuted till the launch of the satellite as part of the algorithm development effort.

Strata are determined by the intersection of a clustering based on the land use / land cover and another based on geographical segmentation. The following Table 4-1 and Figure 4-1 provide the stratification based on the work done until now. A roadmap for future improvements will be discussed later. The classes of the CCI thematic map are grouped into a limited number, whose scattering responses can be assumed comparable. For this purpose, the legend of input CCI thematic map was aggregated as in Table 4-1. Similarly, the geographical segmentation accounts for different climatic areas and environmental conditions and is presented in Figure 4-1.

Table 4-1.Aggregation of the CCI land classes for stratification. The red rows are classes where the SSM algorithm will not be applied (masked classes)

Coverage type	CCI class	New class
No Data	0	0
Cultivated terrain	10, 11, 12, 20	10

Cultivated terrain and natural vegetation	30,40	30
Evergreen forest	50, 70, 71, 72, 90	50
Deciduous forest	60,61,62, 80, 81, 82	60
Forest and natural vegetation	100, 110,160, 170	100
Natural vegetation	120, 121, 122	120
Schabrland, medaw	130, 140	130
Low density vegetation	150, 151, 152, 153	150
Wetlands	180	180
Urban areas	190	190
Bare soil	200, 201, 202	200
Water	210	210
Permanent ice	220	220

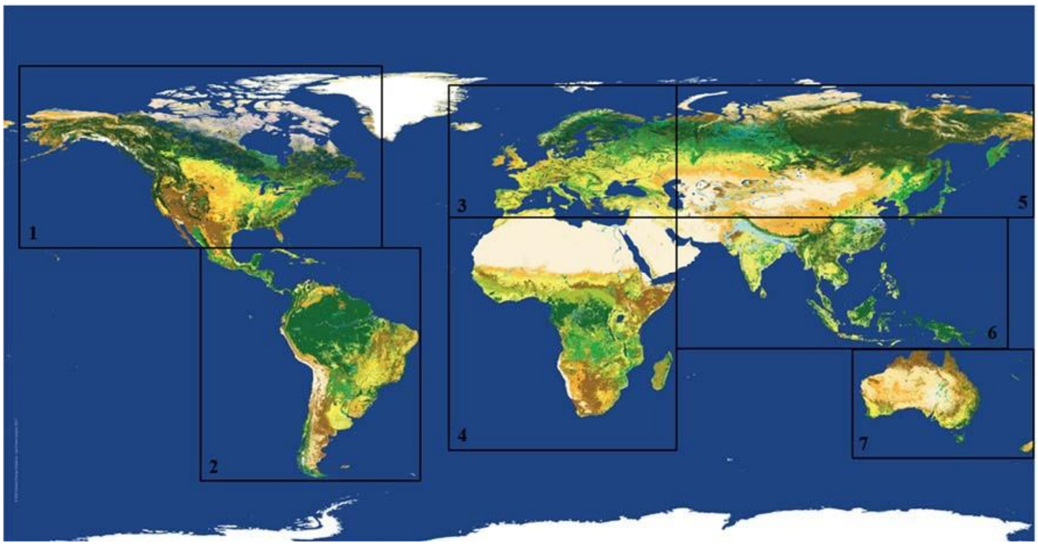


Figure 4-1.Segmentation of the EASE grid based on geographic areas.

The non-linear fitting was based on the simplified eq. (3.14) grouping the two roughness terms into the AFS parameter. An estimate of the unknown parameters in the equation (the AFS and the vegetation σ), both assumed time invariant, was based on an historical dataset, whereas the vegetation optical depth (VOD) was considered time varying and thus collected from other sources. Note that the relationship between VOD and σ is neglected.

As for the present stage, the CyGNSS and SMAP product were used for both purposes. Thus, once collected a suitable sample (one year) of CyGNSS Level1b version 3.0 (effective reflectivity Γ_{eff}) and SMAP L3 (Soil Moisture SM, VWC and VOD) aggregated according to the EASE Grid at 25 km resolution, the dataset was divided with sequential sampling into two equal parts. Moreover, quality flags were applied to filter out bad SMAP (Chan & Dunbar, 2021) and CyGNSS (Ruf et al., 2022)) data. Namely, we considered the following specific quality flags from CyGNSS L1b products: 2, 4, 5, 8, 16, and 17, which are, in order, *S-band transmitter powered up*, *spacecraft attitude error*, *black body DDM*, *DDM is a test pattern*, *direct signal in DDM*, and *low confidence in the GPS EIRP estimate*. As for SMAP we considered the *overall quality* flag. Then the following objective function was minimized by using the first part of dataset for each EASE grid cell to retrieve AFS and σ_0^o.

$$f(x) = \Gamma_{eff}^{CYG} - [e^{-2\tau} \cdot AFS \cdot |R(SM)|^2 + \sigma_0^o(1 - e^{-2\tau})S] \qquad (4.1)$$

$R(SM)$ is computed using the Fresnel reflection coefficient formula. S depends on known geometric quantities. The way to compute the optical depth τ can be different. Using SMAP it is straightforward to use the VOD in the SMAP products. Otherwise, one can use an estimate of the VWC from other sources, including optical data through suitable empirical relationships between optical indices and VWC, but in this case another unknown must be estimated, that is the coefficient b introduced before. As an example of the proposed approach, the following Figure 4-2 provide a map of AFS in each ease grid cell and Table 4-2 reports the values of the vegetation parameter for each strata (intersection of land use and geographical clustering).

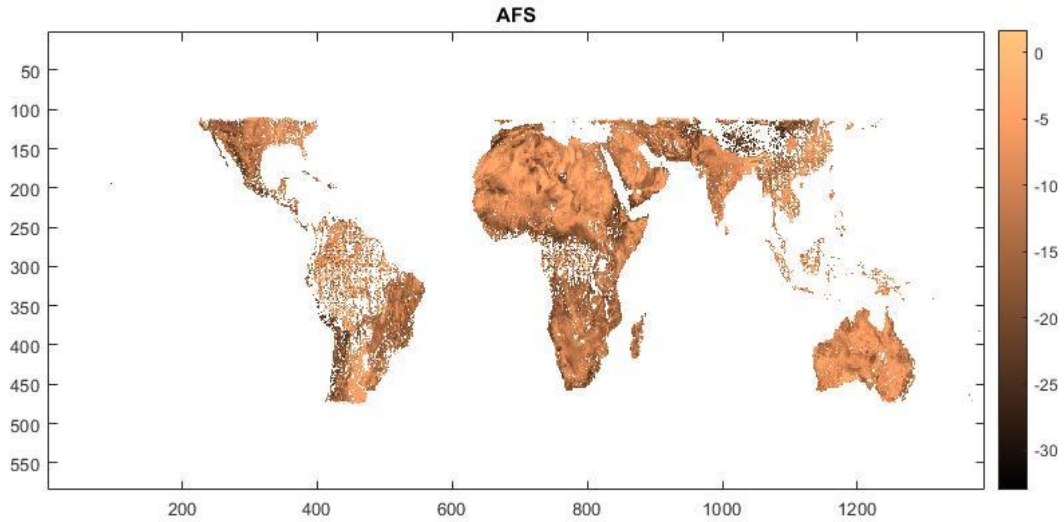


Figure 4-2. Preliminary estimate of the AFS roughness parameter for each EASE grid cell at 25 km resolution based on CyGNSS and SMAP data on 2018 (August to December)

Table 4-2. Preliminary estimates of the σ_0^o parameter for each stratum defined before based on CyGNSS and SMAP aggregated into EASE grid 25km resolution data on the first part of dataset. Note some NaN's (not a number) when the 4-month training set was not covering some strata

	North America	South America	Europe	Africa	Continental Asia	Tropical Asia	Australia
Cultivated terrain	1.0134	1.0166	1.0062	1.0163	1.0095	1.0142	1.0129
Cultivated terrain and natural vegetation	1.0311	1.0178	1.0062	1.0109	1.0147	1.035	1.012
Evergreen forest	1.0251	1.1447	1.0167	1.1001	1.057	1.1237	1.0427
Deciduous forest	1.0236	1.0323	1.0172	1.0143	1.0117	1.0176	1.0262
Forest and natural vegetation	1.031	1.0393	1.0313	1.0116	1.0322	1.0527	1.0196
Natural vegetation	1.0025	1.0093	1.0026	1.0112	1.0036	1.0762	1.0157
Schabrland, medaw	1.0085	1.012	1.009	1.0047	1.0007	1.0000	1.0076
Low density vegetation	NaN	1.0005	1.0023	1.0018	1.0008	1.0028	1.0063
Bare soil	1.0011	1.0004	1.0014	1.0014	1.0013	NaN	1.0033

Figure 4-3 shows the result of the forward model on the first part of the dataset. Although there are points far from the one-to-one perfect matching line, the global result of the forward model shows a fairly good agreement between the best fitted reflectivity and the CyGNSS reflectivity with RMSE in linear units equal to 0.0386, and RMSE in decibel units equal to 2.97. The correlation coefficient is 0.83.

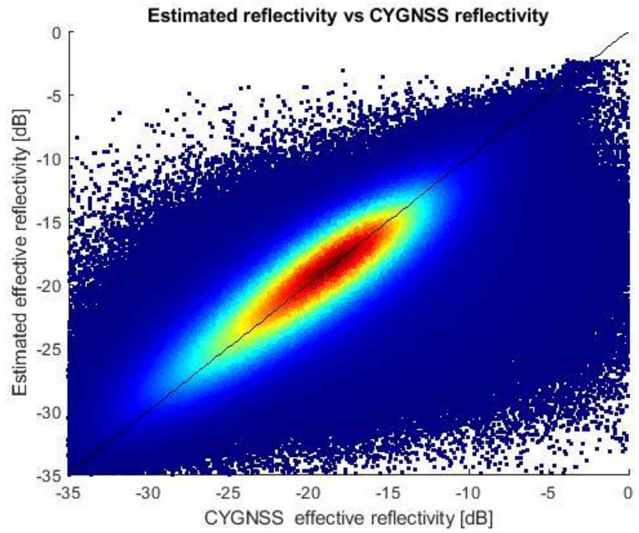


Figure 4-3. Estimated reflectivity and CyGNSS reflectivity on the first part of the dataset.

Then, the following objective function was minimized by using the second part of the dataset for each EASE grid cell to retrieve the coefficients a, b, c, and *d* based on estimations of AFS and σ_0^o from the first part of the dataset.

$$f(x) = SM_{est}^{CYG} - \left[a\left(\Gamma_{eff}^{CYG} - \sigma S\right)_{dB} + b\tau + cAFS_{dB} + d\right] \tag{4.2}$$

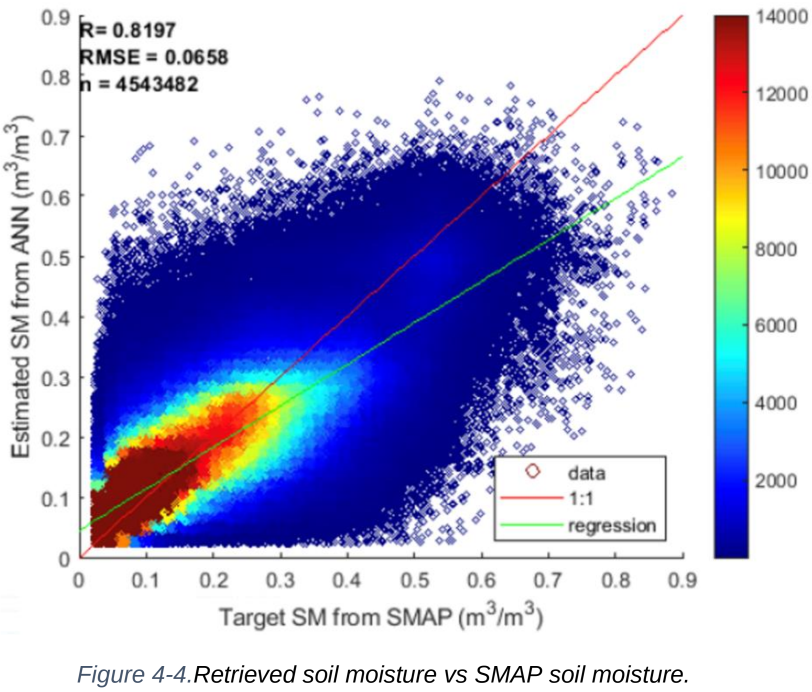


Figure 4-4.Retrieved soil moisture vs SMAP soil moisture.

The trained model (with its retrieved coefficients) was also applied to a dataset spanning another 6 months, from August 1st 2019 to end of January 2020, to assess the performance of the model on an independent set.

The results showed an acceptable agreement between the retrieved soil moisture and SMAP soil moisture, which were RMSE=0.063 and correlation=0.83 for the second part of dataset and RMSE=0.066 and correlation=0.82 for the independent dataset. .Figure 4-4 shows the result of retrieved soil moisture against SMAP reference soil moisture. These performances are within the user requirements specified in the MRD, but better performances have been obtained using the ANN approach developed in the beta release as reported in the following section.

4.1.2 Artificial Neural Network (ANN)

Similar to section 4.1.1, performance assessment of the ANN algorithm network introduced in section 3.6.3 relays on global data taken from CyGNSS and proper auxiliary information. The Cyclone Global Navigation Satellite System (CYGNSS), Soil Moisture Active Passive (SMAP), MODIS, and additional sources like ESA-CCI and a DEM at 1 km resolution are serving as the primary sources. This dataset was resampled and aggregated into the EASE grid V2.0 at 25 km resolution. Moreover, quality flags were applied on SMAP and CyGNSS data. The work was focused on one year, from August 2018 to July 2019. Furthermore, another 6 months of dataset (from August 2019 to the end of January 2020) was used as an independent dataset to assess the performance.

Training and testing of the network was done considering different combinations of inputs, with results summarised in table 4-3. We started just using the CyGNSS variables (denoted as variable* in table 4-3), i.e., Reflectivity (dB), Kurtosis, incidence angle θ, Trailing Edge TE, Coherence index (see Al-Khaldi et al., 2021), SNR (in dB). Then we added static auxiliary data from other sources, like the surface height (DEM), the mean slope (SLOPE) and the rms height (rmsHEIGHT) derived from the DEM in each EASE Grid cell, and the AFS as indicators of topography. Additionally, dynamic auxiliary data were considered, such as the NDWI and NDVI from MODIS, and VOD and VWC from SMAP as indicators of variable vegetation conditions. A comparison of the retrieval performances for different combinations of inputs is reported in Table 4-3. The table shows the importance of considering the auxiliary data, both static and dynamic. As for dynamic data, better performances are obtained gathering it from SMAP products, but MODIS products represent a comparably good source to be used as backup in case SMAP data are not available.

Table 4-3. Performances of the ANN algorithm with different combinations of inputs

Network inputs	Corr coef.	RMSE
CYGNSS variables*	0.48	0.096
CYGNSS variables, SNR (dB), AFS, DEM, SLOPE, rmsHEIGHT	0.77	0.07
CYGNSS variables*, SNR (dB), AFS, DEM, SLOPE, rmsHEIGHT, LCC, AGB, MODIS_NDVI, MODIS_NDWI	0.91	0.045
CYGNSS variables*, SNR (dB), AFS, DEM, SLOPE, rmsHEIGHT, LCC, SMAP_VWC, SMAP_VOD	0.91	0.044

Moreover, these results shows that ANN algorithm outperforms the semi-empirical model performance (see section 4.1.1) in retrieving soil moisture. The ANN algorithm was assessed also on the independent dataset. Figures 4-5 and 4-6 illustrate the performances of the ANN algorithm in two scenarios, with SMAP or with MODIS dynamic auxiliary inputs. These results showed a good agreement between retrieved soil moisture and SMAP soil moisture with RMSE=0.047 m^3/m^3, correlation=0.91 in case of using SMAP and with RMSE=0.048 m^3/m^3, correlation=0.91 in case of using MODIS inputs data.

Another interesting outcome of the pre-launch validation of the ANN algorithm was assessing the performance of the algorithm when there is no dynamic auxiliary dataset available, neither SMAP nor MODIS. The second row of the table 4-3 shows that the performances of the ANN in this situation are still acceptable, so that HydroGNSS data shall be processed in this event and a quality flag in the output product shall be added to indicate lower quality due to lack of dynamic auxiliary data.

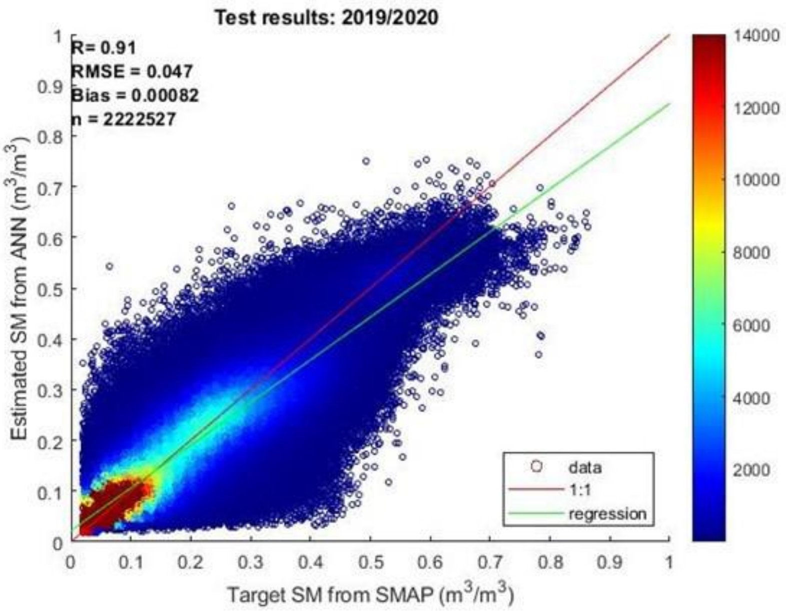


Figure 4-5. Retrieved soil moisture as a function of SMAP reference soil moisture using dynamic auxiliary MODIS data.

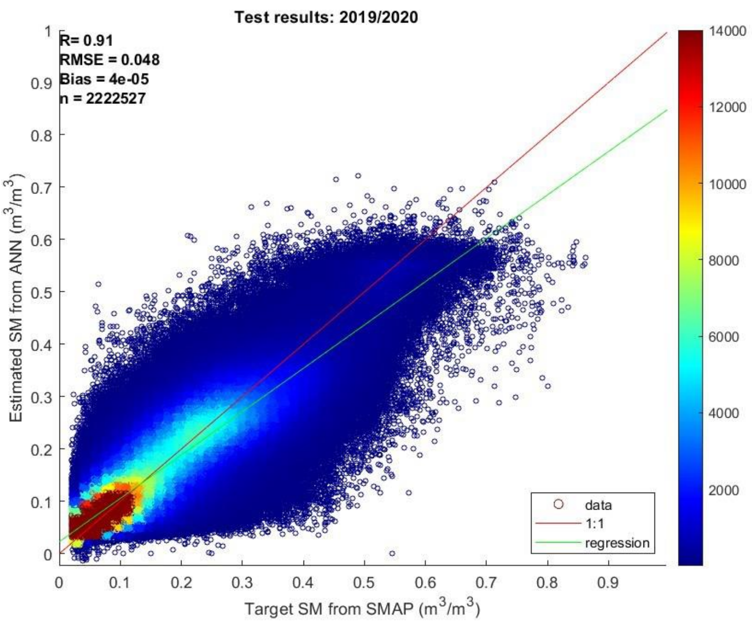


Figure 4-6. Retrieved soil moisture as a function of SMAP reference soil moisture using SMAP dynamic auxiliary data.

4.2 POST LAUNCH VALIDATION

During the HydroGNSS commissioning phase a reliable set of reference SSM measurements from other satellite missions and from ground-based networks or hydrological models must be made available. The data that can be envisaged at the present stage are not always suitable to validate a GNSS-R based product and are sometime not very reliable. The actual baseline for reference data collection is:

- Soil Moisture product from SMAP or SMOS. Their resolution is in the order of 40 km, so they are not suitable to validate GNSS-R based retrievals when attempting to exploit their possible better resolution.

These are however the primary source of reference data for validation of lower resolution HydroGNSS products.

- In situ data from the International Soil Moisture Network. These data are not always reliable, and the spatial distribution of the probes are not always well suited to validate GNSS-R product with a near random spatial distribution. A long data collection can be required to gather a suitable validation set to be assessed at higher resolution. This will be a primary source of reference data, but a careful selection of the networks and quality check of the data must be carried out.

Additional options, that are presently not foreseen yet as no fundings are allocated, could be:

- A ground based reliable network of in-situ instrumentation (soil moisture and soil temperature probes, meteorological data). It could be shared with other ESA validation efforts or promoted in the frame of international cooperations (e.g., sharing validation sites of NASA NISAR mission in the frame of ESA/NASA agreement for ROSE-L). The network should cover an area in the order of 100km, quite flat, and the probes should be spatially distributed to span a range of distances among them from about 5 km to 30 km, thus enabling a test of sensitivity of HydroGNSS to spatial changes of soil moisture at different scales (within the expected range of HydroGNSS resolution).
- Carrying out an airborne experiment collecting data from a microwave imaging radiometer to generate maps of soil moisture.
- Getting output of well calibrated hydrological models over specific sites providing the surface soil moisture as well as soil moisture profiles.

Validation data will be used to compute the performance metrics described in section 4.2.1. Based on those metrics a final decision about the algorithm to be select for operations (physically based or ANN) will be made.

4.2.1 Validation methodology and metrics

The validation task implies a colocation in space and time between HydroGNSS products and reference soil moisture data. Reference works on validation approaches undertaken for SMAP and ASCAT soil moisture products are (Colliander et al., 2017) (Colliander et al., 2022) (Paulik et al., 2014). Colliander at al., 2017 provided details on how SMAP products were matched with reference core site data and can provide a baseline for the matching procedure to be implemented for HydroGNSS. In (Gruber et al., 2020) a wider discussion about issues when validating a soil moisture product is provided and worth to be carefully considered.

According whether the L2G or L3 product is validated, the temporal matching and colocation is different. L2G products are almost instantaneous retrievals and then should be combined with the reference data closest in time. In case reference data have a high sampling rate (e.g., in situ probes) they can be averaged over a small temporal interval around HydroGNSS time (typically one hour) to improve precision. The time difference between reference data and HydroGNSS shall be computed and attached to the collocated pair. It can be a sort of quality index of the matched pair. As for L3 soil moisture maps combining HydroGNSS data within a given time interval, the reference data shall be averaged over the same time frame. It is understood that the validation result may depend on the chosen time resolution.

Regarding spatial colocation, reference data should be posted over the same reference grid of HydroGNSS product. In case of in situ probes, this will be done by averaging the ground measurements within the same grid cell (spatial upscaling). As for reference soil moisture maps from imaging sensors or hydrological models an aggregation (averaging and resampling) or disaggregation (resampling) algorithm shall be applied. Reference can be given to the approach undertaken in (Colliander et al., 2017).

As a certain number of pairs of HydroGNSS soil moisture and reference soil moisture, and their distance in time is produced, a number of graphic products and statistical parameters shall be generated, namely:

- Scatterplot showing the retrieved SM values as a function of reference ones.
- In case of in situ probes with high-rate output, temporal plots of reference measurements with superimposed HydroGNSS single shot SM retrievals
- Comparison of histograms of HydroGNSS SM and reference one over a certain time and spatial span.
- Computation of bulk performance metrics, among which:
  - Root Mean Square Error (RMSE)
  - Unbiased RMSE (uRMSE)
  - Slope of the best fitting line of retrieved as a function of reference soil moisture
  - Pearson Correlation coefficient

4.2.2 Triple colocation

A powerful approach to estimate the error of a source measuring a geophysical parameter (SSM in our case) is the so called "triple colocation". It can be applied when at least three sources of the same target parameter are available, and they can be collocated in space and time. The main idea is that in case the errors of the source are statistical independent it is possible to estimate the standard deviation of the errors of each source starting from the variance and covariance matrices of their outputs.

Detail about triple colocation can be found in the literature, e.g., in (Stoffelen, 1998) (Pierdicca et al., 2015). A brief summary is reported hereafter. Suppose that three measurement systems *X*, *Y*, and *Z* are measuring a true variable θ (*SSM* in this case). Let us assume the following model error for measurements *x*, *y*, *z* provided by the three systems:

$$
\begin{aligned}
x &= s_x(\theta + b_x + \delta x) \\
y &= s_y\left(\theta + b_y + \delta y\right) \\
z &= s_z\left(\theta + b_z + \delta z\right)
\end{aligned}
\tag{4.3}
$$

where δx, δy and δz represent the random observation errors with zero mean, i.e., $<\delta_x>=<\delta_y>=<\delta_z>=0$, and variance $\epsilon_x^2=<\delta_x^2>$, $\epsilon_y^2=<\delta_y^2>$, $\epsilon_z^2=<\delta_z^2>$, while s_x, s_y and s_z are scaling factors, or gains of the systems supposed to have a linear response, and b_x, b_y and b_z account for mean errors different from zero. The *X* system is assumed as the reference, so s_x=1 and b_x=0. The true variable is considered varying randomly, with variance $\sigma^2=<(\theta-<\theta>)^2>=<\tilde{\theta}^2>$, being $\theta=\tilde{\theta}+<\theta>$ with mean value $<\theta>$ not necessarily zero. Usually, the three systems do not represent the same spatial scale of the observed field θ, and thus it is assumed that *X* and *Y* can resolve smaller scales than that resolved by *Z*. Hence the variance common to the smaller scales, i.e., $\sigma_r^2=r^2$ is introduced; it represents, by definition, the correlated part of the representativeness errors of *X* and *Y*. In other words, θ refers to the large-scale features of the observed field, which is measured by *Z*, whereas the small scale features sensed by *X* and *Y* are embedded in the noise terms δ_x and δ_y. Except for the representativeness error, it is assumed that the errors of the different observation systems are not correlated, i.e. $<\delta_x\delta_z>=<\delta_y\delta_z>=0$, and also not correlated with the true random variable, i.e., $<\theta\delta_x>=<\theta\,\delta_z>=<\theta\,\delta_y>=0$. Conversely, because of the representativeness error, it is assumed $<\delta_x\delta_y>=r^2$. With the above assumptions it is possible to derive the unknown error structure by computing some statistical moments of the collocated database. Denoting the correlation coefficients among observations as ρ_{xy}, ρ_{xy}, and ρ_{xy}, it can be demonstrated that the variance of the random errors affecting the three systems can be expressed as:

$$
\begin{aligned}
\varepsilon_x^2 &= \sigma_x^2\left[1-\rho_{xy}\rho_{xz}/\rho_{yz}\right]+\sigma_r^2 \\
\varepsilon_y^2 &= \sigma_x^2\left[\rho_{xz}^2\big/\rho_{yz}^2-\rho_{xy}\rho_{xz}/\rho_{yz}\right]+\sigma_r^2 \\
\varepsilon_z^2 &= \sigma_x^2\left[\left(\rho_{xy}/\rho_{yz}-\sigma_r^2\big/\sigma_x^2\rho_{xz}\right)^2-\rho_{xy}\rho_{xz}/\rho_{yz}\right]+\sigma_r^2
\end{aligned}
\tag{4.4}
$$

The triple colocation can be applied to validate HydroGNSS products by collocating them with SMAP data and output of a hydrological model, provided the latter will be made available in the frame of the Scout program. It is recommended to allocate resources to enable this possibility.

5 ROADMAP FOR FUTURE ALGORITHM IMPROVEMENTS

Cal/val and algorithm development activities generally continue along the entire mission lifetime. Here we list some activities that are foreseen in the frame of the running project and within the end of the Commissioning phase. Additional developments are deemed important for a full exploitation of the mission but will need allocation of additional funding and efforts.

Among the investigations to be done before launch and during Commissioning we identify the following:

- Further extending the calibration/validation of the semi-empirical model and ANN using a larger amount of CyGNSS and SMAP products as well products that are being delivered by the mission during Commissioning.
- Assessing the performances with improved resolution of the final SSM product in EASE grid projection till 9 km. This requires high resolution soil moisture reference values possibly available through the ISMN. It must be considered their quality can be questionable, apart from selected networks.
- Further assessing the reliability of the NDVI and NDWI from MODIS to account for vegetation effects in case SMAP product will not be available in the future.
- Assessing further stratification approaches, especially for what concerns the geographical stratification, better delineating areas with different behaviour. Very fine stratification is proposed in some papers. It must be considered, however, that a very fine geographical stratification generally improves the overall retrieval results but creates discontinuities in the output map between adjacent strata.

Further investigations suggested for future algorithm development:

- Trying different resolutions of the auxiliary data to be combined with HydroGNSS data, especially the DEM derived parameters, in order to optimally match the actual area that is producing the signal. It is well known that the actual resolution of GNSS-R can vary depending on the type of scattering mechanism; than a better correlation with the large-scale roughness and thus the capability to correct its effect, could be found computing the topographic parameters at different horizontal scales.
- Assessing the impact of introducing additional observables as input to the algorithm derived from the whole DDM. In the literature, the waveform leading and trailing edge, the degree of coherent and similar parameters are considered. The Solar Induced Fluorescence (SIF) product available from some satellite missions could be also considered as dynamic auxiliary data instead of NDVI. Integrating additional inputs is generally more feasible in the frame of an ANN data-driven approach.
- As a long-term objective, it can be foreseen the use of Convolutional Neural Networks getting in input the whole DDM to learn about the Earth surface condition around the nominal specular point.
- Implement an ad-hoc experimental campaign best suited to validate GNSS-R retrievals by a suitable spatial distribution of soil moisture probes or by flying an imaging instrument able to estimate soil moisture (e.g., a microwave radiometer). The scope is finally assessing the spatial resolution of the GNSS-R being able to compare with reference SSM data at different spatial scale.

6 CONCLUSIONS

The algorithm for soil moisture retrieval operational processor was described in this document. In the beta version of the L2OP-SSM processor two approaches are implemented: a semiempirical model (the alfa release) and an Artificial Neural Network. Both are based on a stratification of the Earth surface based on geographical location and land use / land cover. As a physical model for the Right-Right polarization is not available in the literature yet, such observable is introduced in a purely empirical way and was experimented only by using simulated data.

The physically based algorithm is being compared to the ANN data-driven approach to make the final selection of the L2OP-SSM operational processor, i.e., the selection of the final algorithm to be used within the PDGS during satellite operations. Additionally, the algorithm coefficients (or the net structure and weights) shall be further tuned on global data to enable processing of HydroGNSS at global scale. The software architecture is designed in such a way that a change in the model and in the stratification does not require to change and recompile the source code.

Presently, the algorithm training on real data was feasible only using CyGNSS and then in the tropical latitude band. During the commissioning phase the HydroGNSS L1b calibrated observables shall be compare first to CyGNSS in order to assess the reliability of the models based on the latter mission. Provided the assessment is satisfactory the validation of the products in the tropical latitude band will be carried out in parallel to the training and validation of the algorithm outside that area.

7 REFERENCES

Al-Khaldi M. M., J. T. Johnson, S. Gleason, E. Loria, A. J. O'Brien and Y. Yi, "An Algorithm for Detecting Coherence in Cyclone Global Navigation Satellite System Mission Level-1 Delay-Doppler Maps," in IEEE

Transactions on Geoscience and Remote Sensing, vol. 59, no. 5, pp. 4454-4463, May 2021, doi: 10.1109/TGRS.2020.3009784."

Camps A., L. Cenci, M. P. Clarizia, D. Comite, F. Costantini, L. Dente, L. Guerriero, A. Mollfulleda, S. Paloscia, H. Park, N. Pierdicca, J. Reynolds, E. Santi, M. Zibri, "Potential of Spaceborne GNSS-R for Land Applications - FINAL REPORT", GLA-FR/01.2019 Issue/Rev. 1.0 / 1.0, September 30, 2019

Camps A., M. Vall-llossera, H. Park, G. Portal, and L. Rossato, "Sensitivity of TDS-1 GNSS-R reflectivity to soil moisture: Global and regional differences and impact of different spatial scales," Remote Sens., vol. 10, no. 11, pp. 1856-1–1856-19, Nov. 2018. doi: 10.3390/rs10111856

Chan, S., & Scott Dunbar, R. (2021). "Soil Moisture Active Passive (SMAP) Mission Level 3 Passive Soil Moisture Product Specification", Document Version 8.0 R18 Release. https://pdms.jpl.nasa.gov/

Chris Ruf et al, 2022. CYGNSS Handbook. Michigan Publishing Services. https://doi.org/10.3998/mpub.12741920.

Christoph Paulik∗, Wouter Dorigo, Wolgang Wagner, Richard Kidd, "Validation of the ASCAT Soil Water Index using in situ data from the International Soil Moisture Network", International Journal of Applied Earth Observation and Geoinformation, 30, (1-8), 2014

Clarizia, M.P. & Pierdicca, Nazzareno & Costantini, Fabiano & Floury, Nicolas. (2019). Analysis of CYGNSS Data for Soil Moisture Retrieval. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. PP. 1-9. 10.1109/JSTARS.2019.2895510

Clarizia, M.P. & Pierdicca, Nazzareno & ostantini, Fabiano & Floury, Nicolas. (2019). Analysis of CYGNSS Data for Soil Moisture Retrieval. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. PP. 1-9. 10.1109/JSTARS.2019.2895510

Colliander A. et al, "Validation of SMAP surface soil moisture products with core validation sites", Remote Sensing of Environment, Volume 19, Pages 215-231, , 15 March 2017.

Colliander A. et al, "Validation of Soil Moisture Data Products From the NASA SMAP Mission", IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, VOL. 15, 2022

Comite Davide, Ticconi Francesca, Dente, Laura, Guerriero Leila, Pierdicca, Nazzareno, "Bistatic Coherent Scattering From Rough Soils With Application to GNSS Reflectometry", IEEE Transactions on Geoscience and Remote Sensing. 10.1109/TGRS.2019.2938442, 2019.

Dente Laura, Leila Guerriero, Davide Comite, Nazzareno Pierdicca, "SpaceBorne GNSS-R Signal Over a Complex Topography: Modeling and Validation". IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 13: 1218-1233 (2020)

Ferrazzoli P., L. Guerriero, N. Pierdicca, R. Rahmoune, "Forest biomass monitoring with GNSS-R: theoretical simulations", Advances in Space Research, Vol. 47, pp. 1823—1832, DOI: 10.1016/j.asr.2010.04.025

Gruber A., G. De Lannoy, C. Albergel, A. Al-Yaari, L. Brocca, J.-C. Calvet, A. Colliander, M. Cosh, W. Crow, W. Dorigo, C. Draper, M. Hirschi, Y. Kerr, A. Konings, W. Lahoz, K. McColl, C. Montzka, J. Muñoz-Sabater, J. Peng, R. Reichle, P. Richaume, C. Rüdiger, T. Scanlon, R. van der Schalie, J.-P. Wigneron, W. Wagner, "Validation practices for satellite soil moisture retrievals: What are (the) errors?", International Journal of Applied Earth Observation and Geoinformation", Remote Sensing of Environment, Volume 244, 111806, July 2020.

Mironov V. L., L. G. Kosolapova, and S. V. Fomin, "Physically and mineralogically based spectroscopic dielectric model for moist soils," IEEE Trans. Geosci. Remote Sens., vol. 47, no. 7, pp. 2059–2070, Jul. 2009.

Pierdicca N. et al., "The Potential of Spaceborne GNSS Reflectometry for Soil Moisture, Biomass, and Freeze–Thaw Monitoring: Summary of a European Space Agency-funded study," in IEEE Geoscience and Remote Sensing Magazine, vol. 10, no. 2, pp. 8-38, June 2022, doi: 10.1109/MGRS.2021.3115448.

Pierdicca N., F. Fascetti, L. Pulvirenti, R. Crapolicchio and J. Muñoz-Sabater, "Quadruple Collocation Analysis for Soil Moisture Product Assessment," in IEEE Geoscience and Remote Sensing Letters, vol. 12, no. 8, pp. 1595-1599, Aug. 2015, doi: 10.1109/LGRS.2015.2414654.

Pierdicca N., L. Guerriero, R. Giusto, M. Brogioni, A. Egido, “SAVERS: AN END-TO-END SIMULATOR OF GNSS REFLECTIONS FROM BARE AND VEGETATED SOILS”, IEEE Transactions on Geoscience and Remote Sensing, Volume: 52 , Issue: 10, DOI: 10.1109/TGRS.2013.2297572, Page(s): 6542 – 6554, October 2014.

Rodríguez-Fernández, N. J., Mialon, A., ermoz, S., Bouvet, A., Richaume, P., Al Bitar, A., Al-Yaari, A., Brandt, M., Kaminski, T., Le Toan, T., Kerr, Y. H., and Wigneron, J.-P.: An evaluation of SMOS L-band vegetation optical depth (L-VOD) data sets: high sensitivity of L-VOD to above-ground biomass in Africa, Biogeosciences, 15, 4627–4645, https://doi.org/10.5194/bg-15-4627-2018, 2018.

Santi Emanuele, Simonetta Paloscia, Simone Pettinato, Giacomo Fontanelli, Maria Paola Clarizia, Davide Comite, Laura Dente, Leila Guerriero, Nazzareno Pierdicca, Nicolas Floury, “Remote Sensing of Forest Biomass Using GNSS Reflectometry”, IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 13: 2351-2368 (2020)

Stoffelen, A. (1998). Toward the true near-surface wind speed: Error modelling and calibration using triple collocation. J. Geophys. Res., 103, 7755–7766.

Wagner W., G. Lemoine, and H. Rott, “A method for estimating soil moisture from ERS Scatterometer and soil data,” Remote Sens. Environ., 70, pp. 191-207, 1999.

Yueh S. H., R. Shah, M. J. Chaubell, A. Hayashi, X. Xu and A. Colliander, "A Semiempirical Modeling of Soil Moisture, Vegetation, and Surface Roughness Impact on CYGNSS Reflectometry Data," in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-17, 2022, Art no. 5800117, doi: 10.1109/TGRS.2020.3035989

Zribi M., M. Foucras, N. Baghdadi, J. Demarty and S. Muddu, "A new reflectivity index for the retrieval of surface soil moisture from radar data," IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens., vol. 14, pp. 818-826, 2021.